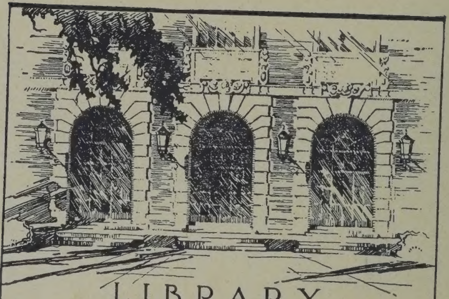


LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

[redacted]

7200-9
140

Proceedings

FIRST ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

DIGITAL COMPUTER
LABORATORY
LIBRARY

November 15-17, 1963
University of Illinois
Allerton House, Monticello, Illinois

Sponsored by the
Department of Electrical Engineering and the
Coordinated Science Laboratory of the
University of Illinois, Urbana, Illinois

OCT 25 1968

PROCEEDINGS
FIRST ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

edited by
J. B. Cruz, Jr.
Conference Chairman

John C. Hofer
Production Editor

DIGITAL COMPUTER
LABORATORY
LIBRARY

Conference held

November 15, 16, and 17, 1963
Allerton House
Monticello, Illinois

Sponsored by the
Department of Electrical Engineering
and the
Coordinated Science Laboratory
of the
University of Illinois
Urbana, Illinois

FOREWORD

The motivation for initiating what is hoped will be an annual series of Allerton Conferences on Systems and Circuit Theory is two fold: (1) We believe that there are not now enough forums for the presentation of new results in a growing dynamic field. (2) The areas described as circuit theory and systems theory are really ONE field and that a unity can be approached by encouraging communications between workers in the two areas.

It is our hope that these Conferences will remain relatively small in attendance, limited to those whose principal interest is theoretical research in systems or circuit theory. There are certainly enough large conferences devoted to a broad coverage of subjects.

Allerton Park of the University of Illinois has many advantages as a location for meetings with relatively small attendance. In addition to its facilities, it has the advantage of isolation. This isolation permits those who participate to live under one roof for the duration of the Conference and to exchange and test ideas without interference. We have been pleased to learn from participants in past Conferences at Allerton that new results stemmed from discussions at Allterton. We hope that this tradition can continue.

We acknowledge the encouragement given by Professor E.C. Jordan, Chairman of the Department of Electrical Engineering at the University of Illinois, and Professor Daniel Alpert, Director of the Coordinated Science Laboratory at Illinois. Arrangements for the meeting and the publication of these PROCEEDINGS were made by Mr. John C. Hofer; registration and other arrangements were made by Miss Rosa Cornell, Mrs. JoAnn Gschwendtner, and Mrs. Mary Barnard. Members of the faculty who participated on the Program Committee included Professors R. B. Ash, W. Mayeda, W. R. Perkins, R. A. Rohrer, S. Seshu, and N. Wax.

J. B. Cruz, Jr.
M. E. Van Valkenburg

FOREWORD

Our motivation for initiating what is hoped will be an annual series of national conferences on Systems and Circuit Theory is two-fold: (1) We believe that there are not now enough forums for the presentation of new results in a growing dynamic field. (2) The areas described as circuit theory and systems theory are really ONE field and that a unity can be approached by encouraging communications between workers in the two areas.

It is our hope that these Conferences will remain relatively small in attendance, limited to those whose principal interest is theoretical research in systems or circuit theory. There are certainly enough large conferences devoted to a broad coverage of subjects.

Allerton Park of the University of Illinois has many advantages as a location for meetings with relatively small attendance. In addition to its facilities, it has the advantage of isolation. This isolation permits those who participate to live under one roof for the duration of the Conference and to exchange and test ideas without interference. We have been pleased to learn from participants in past Conferences at Allerton that new results stemmed from discussions at Allerton. We hope that this tradition will continue.

We acknowledge the encouragement given by Professor E.C. Jordan, Chairman of the Department of Electrical Engineering at the University of Illinois, and Professor Daniel Alpert, Director of the Coordinated Science Laboratory at Illinois. Arrangements for the meeting and the publication of these PROCEEDINGS were made by Mr. John G. Budd; registration and other arrangements were made by Miss Rosa Russell, Mrs. Judith Kalbfleisch, and Mrs. Mary Hubbard. Members of the faculty who participated on the Program Committee included Professors R. B. Ash, W. Mayeda, W. R. Perkins, R. A. Rohrer, S. Seshu, and S. Yau.

J. B. Cruz, Jr.
M. E. Van Valkenburg

TABLE OF CONTENTS

Foreword iii

GENERAL SESSION I Chairman: J.B. Cruz, Jr.

1. Opening Remarks, M. E. Van Valkenburg
2. Basic Ideas in the Cascade Synthesis of Passive N-Ports, D. C. Youla

FILTER DESIGN I Chairman: M. E. Van Valkenburg

1. Approximation of Low-Pass Filters with Little Distortion, F. F. Kuo and M. Karnaugh 1
2. On the Approximation and Synthesis for Tchebycheff Filters with Flat Group Delay Characteristics, T. Iedokoro, Y. Ishizaki and H. Watanabe 16
3. Delay and Attenuation Equalizers with Lossy Reactances, P. R. Geffe 29
4. Linear Phase Band-Pass Filters, R. M. Lerner 36
5. Theoretical Basis of a Digital Computer Program Package for Filter Synthesis, G. Szentirmai 37

TIME-VARYING SYSTEMS Chairman: R. A. Rohrer

1. A Note on Zero-State Stability of Linear Systems, C. A. Desoer and A. J. Thomasian 50
2. Response Functions of Linear Time Varying Systems, J. C. Lindenlaub 53
3. Matrix Representation of Linear Time-Varying Systems, D. C. Lai 63
4. Analysis of Time-Varying Systems: An Analog with Time Invariant Systems, D. D. Weiner and B. Theon 81
5. On Truncation Techniques in the Approximate Analysis of Periodically Time-Varying Nonlinear Networks, I. W. Sandberg 99

NETWORK THEORY I Chairman: B. R. Myers

1. Natural Frequencies of Tunnel Diodes, V. W. Chang and I. T. Frisch 116
2. Transfer Function Synthesis of RC Tunnel-Diode Networks, D. L. Losee and S. K. Mitra 131
3. Analysis of Transients in RC Distributed Parameter Networks, F. A. Lindholm and W. W. Happ 148
4. Design Charts for the Transient Response of RC Distributed Parameter Networks, W. W. Happ and F. A. Lindholm 171
5. Approximating Transfer Functions by Linear Programming, G. A. Matthews, F. Davis, J. C. McKee, R. K. Cavin and G. M. Sibbles 191
6. Digital Computer Analysis of Large Linear Systems, A. H. El-Abiad 205

FILTER DESIGN II Chairman: M. E. Van Valkenburg

1. Design Methods for Sampled Data Filters, J. F. Kaiser 221

2. Fan-out Filters, R. D. Fritzmeyer and C. A. Halijak 237

3. An Extension of Darlington's Semiuniform Predistortion Procedure, G. C. Temes 246

4. On the Existence and Realization of Crystal Ladder Filters, J. D. Schoeffler 259

5. A New Approach in Transfer Function Synthesis, W. H. Chen 275

6. Synthesis of Minimum-Phase Transfer Functions Having Complex Transmission Zeros with Ladder Networks, W. H. Chen 293

NONLINEAR CIRCUITS AND SYSTEMS Chairman: N. Wax

1. Formulation of Liapunov Functions of Nonlinear Systems for Stability Studies, Y. H. Ku 309

2. Nonlinear System Models, T. A. Bickart 320

3. Qualitative Synthesis of Nonlinear Networks, R. Liu and L. Auth 330

4. Limitations of the Classical Volterra Equation Theory in Nonlinear Circuit Problems, B. J. Leon 344

5. Forced Oscillations in Nonlinear Systems with Hysteresis, R. P. de Figueiredo 350

ALLOCATION, SWITCHING, AND GRAPHS Chairman: R. T. Chien

1. Distributed Solution of Network Programming Problems, J. B. Dennis 367

2. Maximal Flow in a Lossy Network, T. Fujisawa 385

3. Optimum Location of Switching Centers and the Absolute Centers and Medians of a Graph, S. L. Hakimi 394

4. Regular Expressions for Linear Sequential Circuits, J. A. Brzozowski 406

5. On Topological Formulae, S. C. Lee 427

BANQUET Toastmaster: B. R. Myers

New Horizons in Circuit and System Theory, J. H. Mulligan, Jr.

GENERAL SESSION II Chairman: J. B. Cruz, Jr.

On a New Characterization of Linear Passive Systems, R. E. Kalman 456

SENSITIVITY Chairman: J. B. Cruz, Jr.

1. Some Results in Multi-Loop Feedback Systems, E. S. Kuh 471

2. Some Bounds on Sensitivity in RLC Networks, M. L. Blostein 488

3. On the Upper Bound and Lower Bound of Excursions of Critical-Frequencies of a Perturbed Network, M. J. Remec 502

4. The Synthesis of Single Variable Systems by State Variable Feedback, B. S. Morgan, Jr. 509

5. The Sensitivities of a Network with Multiple Parameter Variations and Their Minimization, Y. Fu and S. S. Yau 521

N-PORTS AND GRAPHS Chairman: W. Mayeda

1. Degeneracy in N-Port Networks, G. Biorci, P. P. Civalleri, and G. Fiorio 537
2. Characteristics of Nonplanar N-Port Networks, W. H. Kim and T. Falk 562
3. On the Structure of Basic Ambit Sets, H. E. Meadows and B. Liu 569
4. Fundamental Connection Matrices, A. J. Pennington 579
5. Evaluation of Network Determinants via the Key Subgraph, S. D. Bedrosian 589
6. A Method of Network Analysis Using Residual Networks, R. E. Seacat 613

STOCHASTIC PROCESSES AND SYSTEMS Chairman: L. G. Shaw

1. Time-Continuous Information Channels, R. B. Ash 617
2. Optimal Control of Multiplicatively Perturbed Stochastic Systems, W. Gersch and F. Kozin 621
3. A Model for Finite-State Probabilistic Systems, G. D. Bruce and K. S. Fu 632
4. Orthonormal Functions in Adaptive Discrete Systems, D. H. S. Hsu, M. D. Srinath and Y. Fu 652
5. An Optimum Finite Dimensional Signal Representation, R. E. Totty and J. C. Hancock 661

NETWORK THEORY II Chairman: S. Seshu

1. On the Driving-Point Synthesis Problem, R. E. Kalman 662
2. Time-Domain Synthesis of Distributions, A. H. Zemanian 663
3. RLC Driving-Point Synthesis by Distributed Operators, J. B. Murdoch 675
4. A Quaternary Network Algebra, G. E. Sharpe 692
5. On the Use of the Pascal Triangle with Network Functions, K. A. Pullen, Jr. 696

SYSTEM ANALYSIS AND OPTIMIZATION Chairman: W. R. Perkins

1. A Homogeneous Equivalent of Nonhomogeneous Linear State Models and its Application to the Analysis of Continuous and Discrete-State Systems, H. E. Koenig and Y. Tokad 708
2. Optimum Design of Linear Pulse-Width-Modulated Control Systems, T. Meksawan and G. J. Murphy 717
3. Orbit Determination Using Arbitrarily Spaced Data, S. J. Kahne 736
4. Identification and Optimization of Multiparameter Systems Using Correlation Techniques, K. S. Narendra and L. E. McBride, Jr. 740
5. Theoretical Aspects of an Automatic Phase Gain and Root Locus Plotting Device, C. A. Welsh 757
6. Feedback System Analysis by Four-Node Graphs, H. H. Sun and G. C. Reis 764

INDEX OF AUTHORS 779

APPROXIMATION OF LOW-PASS FILTERS WITH LITTLE DISTORTION

F. F. KUO and M. KARNAUGH
Bell Telephone Laboratories, Inc.
Murray Hill, New Jersey

ABSTRACT

A family of gaussian-smoothed rectangle functions is discussed as an aid to bridging the gap between certain system requirements and approximation programs in the synthesis of linear filters. These functions can be used as desired frequency response magnitudes when low-pass filters having

(a) small in-band distortion,

(b) sharp frequency cutoff, and

(c) rapidly damped transients

are desired. They depend upon only two parameters: bandwidth, and a shape parameter ω_o. Therefore, search-evaluation algorithms are relatively easy to apply.

The second part of the paper deals with the role of even numerator polynomials in the rational approximation of a filter amplitude response with linear phase. An IBM 7090 computer program for least squares amplitude approximation is described and a filter approximation example given.

INTRODUCTION

The synthesis of a linear filter often proceeds along the following lines:

(a) Selection of a desired frequency response.

(b) Approximation of the frequency response.

(c) Realization of the approximate response.

Our objectives in this paper are limited to the following: (I) to simplify part of step (a) when a non-distorting low-pass filter amplitude response is desired;

(II) to present a method for the rational approximation of these filter amplitudes with linear phase responses.

In the first part of the paper, a family $G_{\omega_o}(\omega)$ of response magnitude functions is presented where ω_o is a shape parameter. They are useful where low in-band distortion and sharp frequency roll-off must be compromised with a rapidly damped impulse response and low overshoot of the step response. Unlike the rectangle function, these are smoothly varying functions. They can be closely approximated by rational functions when the value of ω_o and the degrees of numerator and denominator are compatible. The frequency, ω, can be arbitrarily rescaled, so there are really two important parameters to be considered by the system designer, ω_o, and α, the frequency scaling parameter.

With appropriately chosen linear phases, the family $G_{\omega_o}(\omega)$ of magnitudes may be used for PCM low-pass filters. Their band-pass and high-pass counterparts also may be used in the design of frequency multiplex systems.

When combined with the approximation algorithm presented below, these functions generate a family of realizable rational response functions. When a figure of merit for filter response can be computed, the best of the approximations so generated can be found by a search-evaluation program.

In the second part of the paper, we present a method for approximating the $G_{\omega_o}(\omega)$ amplitude such that the resulting rational approximation also has the phase linearity of a Bessel filter. The approach is to modify the magnitude response of a Bessel filter without changing its phase response.

I. DEFINITION OF THE MAGNITUDE FUNCTIONS

$$\text{Let } R(\omega) = 1 \quad \text{for} \quad |\omega| \le \omega_o$$

$$R(\omega) = 0 \quad \text{for} \quad |\omega| > \omega_o \tag{1}$$

and let $E(\omega) = \sqrt{2\pi}\,\exp(-\omega^2/2)$. (2)

The convolution of these functions is

$$F(\omega) = R(\omega) * E(\omega) = \sqrt{2\pi}\int_{\omega-\omega_o}^{\omega+\omega_o} \exp(-x^2/2)dx \tag{3}$$

and
$$F(0) = \sqrt{2\pi}\int_{-\omega_o}^{\omega_o} \exp(-x^2/2)dx\ . \tag{4}$$

To normalize for unit gain at $\omega = 0$, let

$$G_{\omega_o}(\omega) = F(\omega)/_{F(0)}\ . \tag{5}$$

This magnitude function results from a gaussian smoothing of a rectangle of half-width ω_o. Its shape is nearly rectangular when $\omega_o >> 1$ and nearly gaussian when $\omega_o << 1$. The effective bandwidth of this function can always be rescaled as desired by a transformation of the form, $\omega' = \alpha\omega$. With this in mind, it seems reasonable to call ω_o a "shape parameter".

Time Response Functions

The impulse response of any linear phase filter having response magnitude $G_{\omega_o}(\omega)$ will differ from the inverse Fourier transform of this function only by a delay. Therefore, the inverse transforms of this family are of some interest.

$$r(t) = \mathcal{F}^{-1}\,[R(\omega)] = \frac{1}{2\pi}\int_{-\omega_o}^{\omega_o} \cos \omega t \, d\omega = \sin \omega_o t/\pi t \qquad (6)$$

$$e(t) = \mathcal{F}^{-1}\,[E(\omega)] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \cos \omega t \exp(-\omega^2/2) d\omega = \exp(-t^2/2) \; . \qquad (7)$$

By the convolution theorem,

$$g(t) = \mathcal{F}^{-1}\left[G_{\omega_o}(\omega)\right] = 2\pi r(t) e(t)/F(0)$$

$$= 2 \sin \omega_o t \exp(-t^2/2)/tF(0) \; . \qquad (8)$$

This is the impulse response of the ideal low-pass filter, modified by gaussian damping. The zeros of this function are not changed by the multiplicative damping.

Frequently, the response h(t) to unit step input u(t) is of some interest.

$$h(t) = u(t) * g(t) = \int_{-\infty}^{t} g(x)dx \qquad (9)$$

From this, it is clear that

$$h(\infty) = \int_{-\infty}^{\infty} g(x)dx = G_{\omega_o}(0) = 1 \qquad (10)$$

But g(x) is an even function, so

$$\int_{-\infty}^{0} g(x)dx = \frac{1}{2} \qquad \text{and}$$

$$h(t) = \frac{1}{2} + \int_0^t g(x)dx = \frac{1}{2} + \frac{2}{F(0)} \int_0^t \frac{\sin \omega_o x}{x} \exp(-x^2/2)dx \ . \tag{11}$$

The maximum overshoot occurs at the end of the principal lobe of g(t), where

$$\omega_o t_m = \pi \ . \tag{12}$$

Therefore, the fractional overshoot is

$$\frac{h(t_m) - h(\infty)}{h(\infty)} = \frac{2}{F(0)} \int_0^{\pi/\omega_o} \frac{\sin \omega_o x}{x} \exp(-x^2/2)dx - \frac{1}{2} \tag{13}$$

Numerical Results

Table I shows the fractional overshoots and Figure 1 shows plots of $G_{\omega_o}(\omega)$ for the values $\omega_o = 1.0, 2.0, 3.0, 4.0,$ and 5.0.

The fractional overshoots were computed according to equation (13), using a 400-point numerical integration.

The corresponding response magnitudes $G_{\omega_o}(\omega)$ were computed as in equations (3),(4),(5), using a 300-point integration for each value of ω.

These computations were done from $\omega = 0$ to $\omega = \omega_o + 3$ at intervals $\Delta\omega = (\omega_o+3)/100$. To conveniently fit these curves on the same coordinates, the transformation

$$\omega' = \frac{5}{\omega_o+3}\,\omega$$

was used. These results are plotted in Figure 1.

II. AMPLITUDE APPROXIMATION WITH LINEAR PHASE

In filter approximation the problem is to approximate, over a given range of frequencies, a given amplitude and phase response by a rational function. A "simplification" of this problem occurs in pulse transmission applications where the phase response is required to be linear over the bandwidth of the system. Storch* has shown that a system function with a constant numerator and a Bessel polynomial denominator has an approximately linear phase characteristic. The amplitude response corresponding to such a system function is low-pass. The question now arises: Can we modify the amplitude response of the Bessel filter to approximate any given low-pass characteristic and still maintain the linearity of the phase response?

The answer to the question requires a method of "phaseless" correction. It is easy to see that a quad of zeros shown in Fig. 2a does not affect the phase at any point along the jω axis. The pair of zeros on the real axis in Fig. 2b changes the phase by 180° at every point along the jω axis. The pair of zeros located at $\pm j\omega_1$ in Fig. 2c does not affect the phase in the range $|\omega| < |\omega_1|$ and changes the phase by 180° for $|\omega| > |\omega_1|$. We see that the addition of zeros as in Figs. 2a, b and c does not change the linearity of the phase response. The only difficulty might arise in the case of jω axis zeros where the phase changes by 180° at every pair of zeros. However, for low-pass filter approximation, all the jω zeros are placed in the stop band where the transmission is negligible because the amplitude response is small. Thomson* has shown that an approximately symmetrical impulse response can be obtained if one uses a Bessel polynomial denominator with a numerator whose zeros are on the $j\omega$ axis. The same thing holds true for more general even polynomial numerators which suppress the out-of-band frequencies (i.e., those frequencies for which phase linearity fails most severely).

---

*Leo Storch, "Synthesis of Constant Time Delay Ladder Networks Using Bessel Polynomials", Proc. IRE, 42, pp. 1666-1676, November, 1954.

The zeros in Fig. 2 do affect the amplitude response, however. We thus have the following problem to solve. Given a set of arbitrary amplitude data points and a rational function whose denominator is prespecified, choose the numerator (an even polynomial with real coefficients) such that the amplitude response of the rational function is a least squares approximant of the specified amplitude data.

We can choose the denominator to be a Bessel polynomial scaled in frequency such that the linear part of the phase response covers at least the bandwidth of the system. Next we specify the degree of the numerator polynomial to account for the asymptotic roll-off. We then determine the numerator polynomial according to the method outlined below. Suppose $|G(j\omega_i)|$ is the specified amplitude data over the set of data points ω_i. Let $D(s)$ be the denominator polynomial. We are required to find the coefficients, a_k, of the even numerator polynomial,

$$N(s) = \sum_{k=0}^{n} a_k s^{2k} .$$

---

*W. E. Thomson, "The Synthesis of a Network to have Sine-Squared Impulse Response", Vol. 99, Part III, Proc. I.E.E., Nov. 1952, p. 373.

To determine the a_k, we perform a least squares minimization on the sum,

$$E(\omega_i) = \sum_i \left[\, | G(j\omega_i) | - \frac{\sum_k^n a_k (j\omega_i)^{2k}}{| D(j\omega_i) |} \right]^2 \qquad (14)$$

Differentiating $E(\omega_i)$ with respect to a_ℓ, we obtain

$$\frac{dE(\omega_i)}{da_\ell} = -2 \sum_i \left[\, | G(j\omega_i) | - \frac{\sum_{k=0}^n a_k (j\omega_i)^{2k}}{| D(j\omega_i) |} \right] \frac{(j\omega_i)^{2\ell}}{| D(j\omega_i) |}, \; \ell = 0,1,2,..,n \qquad (15)$$

Setting (15) equal to zero, we have the following set of n+1 linear equations in a_k,

$$\sum_i \frac{| G(j\omega_i) | \, x_i^{\ell}}{| D(j\omega_i) |} = \sum_{k=0}^{n} \sum_i \frac{a_k x_i^{(k+\ell)}}{| D(j\omega_i) |^2} \qquad (16)$$

where $x_i = -\omega_i^2$. We then solve the set of simultaneous equations in (16) for a_k by the computer program described below.

COMPUTER PROGRAM AND NUMERICAL EXAMPLE

A Fortran program has been written to evaluate the numerator coefficients, a_k, from the set of n equations in (16). This program requires only a subroutine for solving simultaneous equations and an amplitude and phase subroutine. In addition, a root-finding subroutine is desirable, though not necessary. The program requires as input information the set of amplitude data, $| G(j\omega_i) |$, the data points ω_i, the denominator polynomial D(s) given either in factored or un-factored form, and the degree of the numerator polynomial.

The program computes the numerator coefficients a_k and then finds the amplitude and phase of the rational function N(s)/D(s) at the data points, ω_i. It computes the difference between the calculated amplitude and the specified amplitude at each data point and also calculates the sum of the squared differences over the data points. For convenience, the program also calculates the deviation of the phase response from a linear phase obtained by straight-line extrapolation of the phases of the first two data points.

Now, consider a numerical example. The amplitude response to be matched is an $\omega_o = 4$ magnitude shown by the solid curve in Fig. 3. We begin with an all-pole 12th order Bessel filter whose magnitude is given by the dashed-dotted curve in Fig. 3. To the Bessel polynomial denominator we add six zeros such that the resulting magnitude is a least squares approximation of the $\omega_o = 4$ magnitude. The calculations for the sixth degree numerator in Figs. 4 and 5 show that there are two pairs of imaginary axis zeros at

$$\omega = \pm 4.459$$

$$\omega = \pm 14.729$$

and a pair of real axis zeros at

$$s = \pm 1.917$$

The real axis zeros control the shaping of the magnitude curve around $\omega = 0$ and the imaginary zeros perform similar duties at the high end of the curve (above $\omega = 3$). The magnitude of the approximant is shown by the dashed curve in Fig. 3. To obtain a better least squares fit, one must use higher order numerators, which in turn would lead to a less sharp asymptotic roll-off in the stop band and may not be desirable.

The program contains a number of useful features. The first is a provision for prespecifying certain zeros. The program first modifies the magnitude curve to be approximated by an amount equal to the magnitude response of the prespecified zeros and then approximates the modified magnitude with the least squares routine. A case in which the use of prespecified zeros is important is in the design of a class of low-pass filters useful in PCM. For PCM decoders it is useful to have a transmission zero at the sampling frequency. Suppose it is desired to have the magnitude 6 db down at one half the sampling frequency. If we were to use the $\omega_o = 4$ curve for the amplitude response of the PCM filter, the 6 db point is approximately at $\omega = 2.9$. Thus the prespecified zeros must be placed at $\omega = \pm 5.8$. Then, if we calculate an additional four zeros using the least squares routine, these zeros fall at

$$\omega = \pm 5.061$$

$$s = \pm 1.805$$

The least squares difference for this approximation is 0.00734 compared to a difference of .00487 for the case of a 6th degree numerator with no zeros prespecified.

An additional feature of the least squares program is a provision of a frequency weighting function so that certain data points can be weighted more heavily, as desired.

From the calculated results in Fig. 4, we note that there are two pairs of zeros on the $j\omega$ axis at $\omega = \pm 4.45870$

and $\omega = \pm 14.7291$. Between these zeros, the phase runoff is more than 180° due to a phase reversal. However, the amplitude is so far down between these frequencies that the phase reversal makes very little difference in transmission. In applying this method to low-pass filter approximation, we can quite safely assume that no $j\omega$ axis zeros will be placed in the pass band.

CONCLUSION

The selection of a frequency response function can often be reduced to a simple search-evaluation procedure by searching over a very restricted but suitably chosen space. For the case of low-pass filters having little distortion, the family $G_{\omega_o}(\omega)$ appears to be suitably chosen because both the time and frequency functions have approximately gaussian decay.

In conjunction with the selection of the amplitude response function, a least squares amplitude approximation procedure is also presented. The procedure is based upon a simple method for amplitude equalization of low-pass filters by the addition of zeros to all-pole filter transfer functions. The method enables a designer to obtain a transfer function with an arbitrary low-pass amplitude characteristic and linear phase. In contrast to all-pass phase equalizers, the amplitude equalizer cannot be designed as a separate structure to be placed in tandem with the original filter. This is because it is an all-zero function. Instead, the entire filter-equalizer must be redesigned from the transfer function obtained through the method described. The zeros in the right half plane may require the use of lattice networks and/or ideal transformers.

TABLE I

ω_o	Fractional Overshoot
1.0	.00007
2.0	.01193
2.5	.02614
3.0	.04000
3.5	.05105
4.0	.05913
5.0	.06924

FIGURE CAPTIONS

Fig. 1: $G_{\omega_o}(\omega)$ vs. ω for different values of ω_o

Fig. 2: Zero Configurations for "Phaseless" Magnitude Correction

Fig. 3: Amplitude Responses of the Different Filters

Fig. 4: IBM Printout of Numerical Example, Part I

Fig. 5: IBM Printout of Numerical Example, Part II

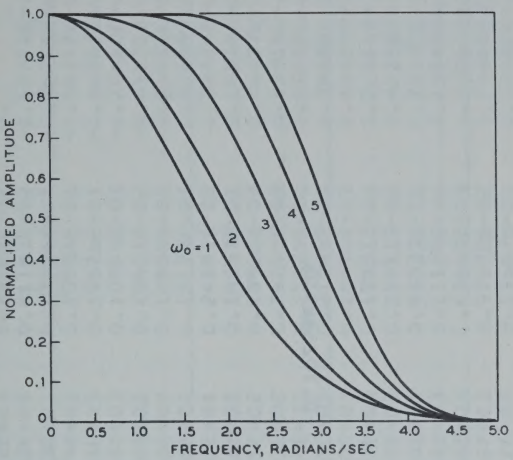
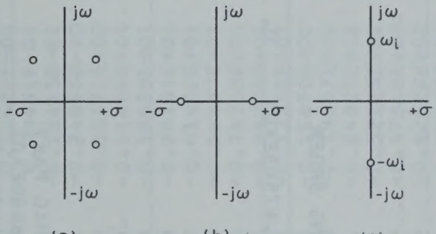


FIG. 1


jω
-σ
+σ
-jω
jω
-σ
+σ
-jω
jω
ω_i
-σ
+σ
$-\omega_i$
-jω

(a) (b) (c)

FIG. 2

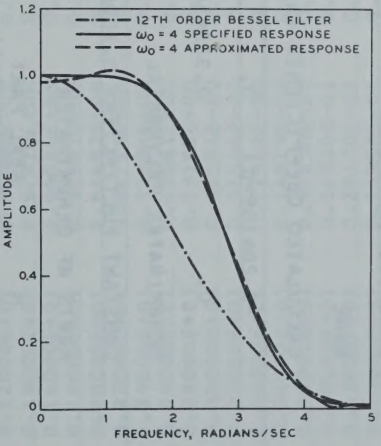


FIG. 3

```
NUMERATØR PØLYNØMIAL

CØNSTANT MULTIPLIER=-0.20410E+03

CALCULATED ZERØS

      N          REAL PART                      IMAG PART
      1       0.19168960E+01        0.
      2      -0.19168960E+01        0.
      3       0.                    0.44587024E+01
      4      -0.                   -0.44587024E+01
      5       0.                    0.14729135E+02
      6       0.                   -0.14729135E+02

PRESPECIFIED ZERØS

NØNE

CALCULATED CØEFFICIENTS IN INCREASING ØRDER

     -0.20410E+03      0.              -0.47585E+05     0.              -0.70264E+06

      0.               0.32345E+07

DENØMINATØR PØLYNØMIAL

CØNSTANT MULTIPLIER= 0.10000E+01

RØØTS ØF DENØMINATØR

      N          REAL PART                     IMAG PART
      1     -0.25426920E+01       -0.23736690E+01
      2     -0.25426920E+01        0.23736690E+01
      3     -0.31743920E+01       -0.33372850E-00
      4     -0.31743920E+01        0.33372850E-00
      5     -0.28713769E+01       -0.16808309E+01
      6     -0.28713769E+01        0.16808309E+01
      7     -0.12857769E+01       -0.38939620E+01
      8     -0.12857769E+01        0.38939620E+01
      9     -0.30758770E+01       -0.10034849E+01
     10     -0.30758770E+01        0.10034849E+01
     11     -0.20498849E+01       -0.30972700E+01
     12     -0.20498849E+01        0.30972700E+01
```

FIG. 4

```
FREQUENCY RESPØNSE

  FREQUENCY      ØRIGINAL       CALCULATED     CALCULATED     CALCULATED     AMPLITUDE      PHASE RUNØFF
    W(I)         AMPLITUDE      AMPLITUDE      LØSS IN DB     PHASE DEGS     DIFFERENCE      IN DEGREES
 0.20000E-00    0.99995E+00    0.97874E+00   -0.18666E-00    0.33021E+03    0.21211E-01    0.
 0.40000E-00    0.99977E+00    0.98612E+00   -0.12139E-00    0.30041E+03    0.13648E-01    0.
 0.60000E+00    0.99923E+00    0.99626E+00   -0.32550E-01    0.27062E+03    0.29708E-02    0.
 0.80000E+00    0.99807E+00    0.10062E+01    0.53588E-01    0.24082E+03    0.81183E-02   -0.57220E-05
 1.00000E+00    0.99521E+00    0.10123E+01    0.10622E-00    0.21103E+03    0.17094E-01   -0.76294E-05
 0.12000E+01    0.98988E+00    0.10108E+01    0.93431E-01    0.18124E+03    0.20934E-01   -0.38147E-05
 0.14000E+01    0.97870E+00    0.99823E+00   -0.15415E-01    0.15144E+03    0.19526E-01   -0.38147E-05
 0.16000E+01    0.96084E+00    0.97179E+00   -0.24855E-00    0.12165E+03    0.10950E-01   -0.28610E-05
 0.18000E+01    0.92881E+00    0.92985E+00   -0.63171E+00    0.91856E+02    0.10431E-02   -0.85831E-05
 0.20000E+01    0.88497E+00    0.87209E+00   -0.11888E+01    0.62062E+02    0.12880E-01   -0.85831E-05
 0.22000E+01    0.81774E+00    0.79957E+00   -0.19429E+01    0.32268E+02    0.18173E-01    0.42439E-04
 0.24000E+01    0.73894E+00    0.71465E+00   -0.29181E+01    0.24748E+01    0.24291E-01    0.32425E-03
 0.26000E+01    0.63561E+00    0.62076E+00   -0.41415E+01   -0.27318E+02    0.14847E-01    0.18616E-02
 0.28000E+01    0.53191E+00    0.52205E+00   -0.56457E+01   -0.57105E+02    0.98562E-02    0.86327E-02
 0.30000E+01    0.41556E-00    0.42298E-00   -0.74735E+01   -0.86873E+02    0.74249E-02    0.33968E-01
 0.32000E+01    0.31564E-00    0.32797E-00   -0.96834E+01   -0.11659E+03    0.12325E-01    0.11535E-00
 0.34000E+01    0.21969E-00    0.24103E-00   -0.12358E+02   -0.14615E+03    0.21344E-01    0.34190E-00
 0.36000E+01    0.14919E-00    0.16559E-00   -0.15620E+02   -0.17540E+03    0.16396E-01    0.89128E+00
 0.38000E+01    0.91230E-01    0.10403E-00   -0.19657E+02   -0.20403E+03    0.12801E-01    0.20561E+01
 0.40000E+01    0.54810E-01    0.57351E-01   -0.24829E+02   -0.23165E+03    0.25411E-02    0.42268E+01
 0.42000E+01    0.29170E-01    0.24817E-01   -0.32105E+02   -0.25786E+03    0.43533E-02    0.78139E+01
 0.44000E+01    0.15390E-01    0.42104E-02   -0.47513E+02   -0.28232E+03    0.11180E-01    0.13145E+02
 0.46000E+01    0.70000E-02    0.74366E-02   -0.42572E+02   -0.12486E+03    0.43663E-03    0.20040E+03
 0.48000E+01    0.32700E-02    0.13027E-01   -0.37703E+02   -0.14546E+03    0.97570E-02    0.20959E+03
 0.50000E+01    0.12900E-02    0.14897E-01   -0.36538E+02   -0.16422E+03    0.13607E-01    0.22062E+03

 LEAST SQUARES DIFFERENCE=  0.4873584E-02
```

FIG. 5

ON THE APPROXIMATION AND SYNTHESIS FOR TCHEBYCHEFF FILTERS WITH FLAT GROUP DELAY CHARACTERISTICS

T. IEDOKORO, Y. ISHIZAKI and H. WATANABE
Nippon Electric Company Ltd.
Kawasaki City, Japan

Summary

In this paper, a method for designing Tchebycheff filters having flat group delay characteristics in the main part of the pass band is presented. From the general approximation theory, a method for obtaining the transfer function of Tchebycheff filters with complex transmission zeros is derived, which are so determined that the group delay characteristics may be flat in the main part of the pass band.

A new theory of cascade synthesis which is an extension of "Belevitch's cycle" and an extended "Bartlett theorem" are introduced. By the application of these results, Tchebycheff filters with flat group delay can be realized. A result of using the suggested method in designing a Tchebycheff filter of 17th degree with flat group delay is included.

1 Introduction

The design of filters with Tchebycheff performance in their pass bands has been exhaustively studied by many authors(1). As for the filters used in transmission systems, the group delay characteristics in the pass band has played a more and more important role as well as attenuations. Especially on the filters in both compatible use in telegraph and data transmission, Tchebycheff attenuation characteristics in the pass band are required for speech transmissions and at the same time, the flat group delay performance in the main part of the pass band is also required for wave form transmissions. Many attempts to design filters with flat delay, including classical Bode's ideal filter, have been made, but none of these are Tchebycheff filters.* One of the authors has already suggested a method for improving the delay characteristics of Tchebycheff filters by introducing pertinent complex transmission zeros.(3)

The purpose of this paper is to present the method of filter design which has Tchebycheff performance as well as approximately flat group delay in the pass band. This method consists of approximation and synthesis of reactance networks with complex transmission zeros.

In the first part of this paper, a simple but useful method of determining real and/or complex attenuation poles (transmission zeros), so that the overall transmission characteristics may satisfy the requirements, is described. Fine approximation is, of course, made by iterative techniques using a digital computer.

---

* Recently, the authors found that a similar work had also been suggested independently by J.E. Colin of S.A.T.(2). His method essentially the same as ours but he had not described the approximation method of the delay and synthesis theorem of networks with complex valued elements.

The second part of the paper introduces the main results of the theory of networks with complex valued elements which was developed by one of the authors$^{(5)}$ as an extension of Brune's and Belevitch's positive function synthesis. By application of these results, the reactance network with complex transmission zeros can be easily realized.$^{(6)}$

Lastly, a practical design example is described in which a filter of the 17th order was designed along the suggested method and computed by a digital computer.

2 Group Delay Characteristics of Tchebycheff Filters

A general theory of approximation of filters having Techebycheff performance in the pass band has been developed by the last named author of this paper.$^{(4)}$ According to this approximation theory we can design almost any type of Tchebycheff filter, including the band pass filter with a minimum number of inductors. It is quite useful to know the delay characteristics of such Tchebycheff filters before performing any actual calculation. Fortunately, this is practically accomplished by the fact that the delay characteristics of our Techebycheff filter have strong connections with the function $\psi(\lambda)$ defined in the reference (4).

According to this approximation theory, a generalized characteristic function $\phi(\lambda)$ of a filter which exhibits Tchebycheff performance in the pass band is given by

$$\phi(\lambda) = H \cosh \mathcal{W}(\lambda). \tag{1}$$

Where $\mathcal{W}(\lambda)$ is a kind of transcendental function which is called "ideal transmission function" in reference (4), and is expressed as follows:

$$\mathcal{W}(\lambda) = \int_{\phi_o}^{\lambda} \psi(\lambda)\cdot d\lambda \tag{2}$$

$$\psi(\lambda) = \left[2\gamma_\infty + \sum_{\nu=1}^{h} \frac{2\gamma_{-\nu'}\cdot Q_{-\nu'}\cdot\sqrt{Q_{-\nu'}^2+1}}{Q_{-\nu'}^2 - \lambda^2} + \sum_{\nu=1} \frac{2\gamma_\nu \cdot Q_\nu \cdot \sqrt{Q_\nu^2 - 1}}{Q_\nu^2 + \lambda^2} \right.$$
$$\left. + \sum_{i=1}^{g} 2\gamma_i \left\{ \frac{\lambda_i \cdot \sqrt{\lambda_i^2 + 1}}{\lambda_i^2 - \lambda^2} + \frac{\lambda_i^* \cdot \sqrt{\lambda_i^{*2} + 1}}{\lambda_i^{*2} - \lambda^2} \right\} \right] \frac{1}{\sqrt{1+\lambda^2}}, \tag{3}$$

where

$Q_{-\nu'}$; attenuation poles on the real axis

Q_ν ; attenuation poles on the imaginary axis

$\lambda_i = d_i + jQ_i$; attenuation poles on the complex domain

$\lambda_i^* = d_i - jQ_i$;

$\gamma_\infty, \gamma_{-\nu'}, \gamma_\nu$ and γ_i; residues of $\psi(\lambda)\cdot d\lambda$.

The operating transfer factor $S(\lambda)$ can be found by factorizing the following fundamental equation:

$$1 + \phi^2(\lambda) = S(\lambda)\cdot S(-\lambda). \tag{4}$$

As described in the previous paper by the authors$^{(4)(7)}$, any band pass filter can also be designed from the same equation (4) by means of a linear transformation of the squared variable between the λ- and the p-plane,

$$\lambda^2 = \frac{p^2 + k'^2}{1 - k'^2}, \tag{5}$$

where k' is the specific band ratio. Therefore, without loss of generality, let us consider the $\phi(\lambda)$ and $S(\lambda)$ on the λ plane only, and for the convenience of description suppose $S(\lambda)$ to be a rational function of λ.*

Denote α as a real constant, then equation (1) may be written as

$$\phi(\lambda) = \frac{\cosh \mathcal{W}(\lambda)}{\sinh \alpha}. \tag{6}$$

Equation (4) becomes

$$S(\lambda)\cdot S(-\lambda) = \frac{\cosh\,[\mathcal{W}(\lambda) + \alpha]\cdot\cosh\,[\mathcal{W}(\lambda) - \alpha]}{\sinh^2\alpha}. \tag{7}$$

Now, define a new function of λ as

$$\mathcal{W}^*(\lambda) = \cosh^{-1}\left\{S(\lambda)\cdot\sinh\alpha\right\} - \alpha, \tag{8}$$

then we have

$$S(\lambda) = \frac{\cosh\left\{\mathcal{W}^*(\lambda) + \alpha\right\}}{\sinh\alpha}. \tag{9}$$

From equations (7) and (9), we obtain a relation

$$\frac{\cosh\left\{\mathcal{W}^*(\lambda) + \alpha\right\}}{\cosh\left\{\mathcal{W}(\lambda) + \alpha\right\}} = \frac{\cosh\left\{\mathcal{W}(\lambda) - \alpha\right\}}{\cosh\left\{\mathcal{W}^*(-\lambda) + \alpha\right\}}. \tag{10}$$

At this stage, our problem is to find the difference between $\mathcal{W}(\lambda)$ and $\mathcal{W}^*(\lambda)$ on the entire λ plane. Since $\mathcal{W}(\lambda)$ is a transcendental function, divide the Riemann surface determined by $\psi(\lambda)\cdot d\lambda$ into two parts, R_I and R_{II} by a branch cut $[-j, +j]$, then $\mathcal{W}(\lambda)$ becomes a one-valued function on R_I. So let us consider, hereafter, such one-valued $\mathcal{W}(\lambda)$ on R_I.

Now, let poles Q_ν of $\mathcal{W}(\lambda)$ be also poles of both $\mathcal{W}^*(\lambda)$ and $\mathcal{W}^*(-\lambda)$. This means, poles of $S(\lambda)$ and $\phi(\lambda)$ are identical and located symmetrically with respect to both the real and imaginary axis of the λ-plane.

---

* Naturally $S(\lambda)$ may, in general, be an irrational function for which the same reasoning is still valid.

In the neighbourhood of the zeros of $S(\lambda)$, we have

$$\mathcal{W}^*(\lambda) = j(n + \frac{1}{2})\pi - \alpha, \tag{11}$$

$$\mathcal{W}(\lambda) = \begin{cases} j(n' + \frac{1}{2})\ \pi - \alpha, \\ \text{or} \\ j(n' + \frac{1}{2})\ \pi + \alpha. \end{cases} \tag{12}$$

Since $\mathcal{W}^*(\lambda)$ and $\mathcal{W}(\lambda)$ have the same poles, the integer number n' must be equal to n. Therefore, in this neighbourhood the difference between $\mathcal{W}^*$ and $\mathcal{W}$ becomes

$$\mathcal{E}(\lambda) = \mathcal{W}^*(\lambda) - \mathcal{W}(\lambda) = \begin{cases} 0 \\ \text{or} \\ -2\alpha \end{cases}$$

and

$$\mathcal{W}^*(-\lambda) - \mathcal{W}(\lambda) = \begin{cases} -2\alpha \\ \text{or} \\ 0 \end{cases} \tag{13}$$

In the pass band, $\mathcal{W}(\lambda)$ becomes $j\,\mathcal{V}(\Omega)$ which is a monotonic function of Ω. At the particular point $\lambda = j\Omega_k$ where $\mathcal{V}(\Omega_k)$ has the value of $(k + \frac{1}{2})\pi$ (k is integer number), equation (10) becomes

$$\frac{\sinh\ (\alpha + \mathcal{E})}{\sinh\ \alpha} = \frac{\sinh\ \alpha}{\sinh\ (\alpha + \bar{\mathcal{E}})} \tag{14}$$

where

$$\mathcal{E} = \mathcal{E}(j\Omega_k) = \mathcal{W}^*(j\Omega_k) - \mathcal{W}(j\Omega_k)$$

$$\bar{\mathcal{E}} = \mathcal{E}(-j\Omega_k).$$

From the preceding discussion, we can realize that

$$0 \geq R_e\ \mathcal{E}(j\Omega_k) > -2\ \alpha.$$

The magnitude of $\mathcal{E}$ in the pass band, unfortunately, is not exactly determined by equation (14) only, but from numerical calculations on various filters we can conclude that the magnitude of $\mathcal{E}$ is considerably small compared with $\mathcal{V}(\Omega)$ in the main part of the pass band, although it becomes large at the end of pass band. As an example, the magnitude of $\mathcal{E}$ for a conventional filter is indicated in the figure 1.

Using such new function as $\mathcal{W}^*(\lambda)$, the group delay of a Tchebycheff filter in the pass band, can be easily estimated by the following equation:

$$\tau = R_e \left\{ \frac{\frac{d}{d\lambda} S(\lambda)}{S(\lambda)} \right\}_{\lambda = j\Omega} = R_e \left\{ \tanh \left\{ \mathscr{W}^*(\lambda) + a \right\} \cdot \frac{d}{d\lambda} \mathscr{W}^*(\lambda) \right\}_{\lambda = j\Omega.} \tag{15}$$

If we neglect $\mathcal{E}$ in the pass band, then τ can be expressed approximately as

$$\tau (\Omega) \simeq X(\Omega) \cdot \tau_o (\Omega) , \tag{16}$$

where

$$X(\Omega) = \frac{\sinh \left\{ \mathrm{Re}\, \mathscr{W}^* + a \right\} \cosh \left\{ \mathrm{Re}\, \mathscr{W}^* + a \right\}}{\sinh^2 a \cdot \left| S(j\Omega) \right|^2} ,$$

and

$$\tau_o (\Omega) = \frac{d}{d\Omega} \mathcal{U} (\Omega) .$$

Since $X(\Omega)$ is almost a constant in the main part of the pass band, τ is approximated by $\tau_o(\Omega)$, which can be written from equation (3) as follows:

$$\tau_o(\Omega) = \left[2\gamma_\infty + \sum_{\nu=1}^{\ell} \frac{2\gamma_{-\nu'} \cdot Q_{-\nu'} \sqrt{Q_{-\nu'}^2 + 1}}{Q_{-\nu'}^2 + \Omega^2} + \sum_{\nu=1}^{h} \frac{2\gamma_\nu \cdot Q_\nu \sqrt{Q_\nu^2 - 1}}{Q_\nu^2 - \Omega^2} \right.$$

$$\left. + \sum_{i=1}^{g} 2\gamma_i \frac{D_i + E_i \cdot \Omega^2}{\Omega^4 - B_i \cdot \Omega^2 + C_i} \right] \frac{1}{\sqrt{1 - \Omega^2}} \tag{17}$$

where

$$B_i = 2(Q_i^{\,2} - d_i^{\,2})$$

$$C_i = (Q_i^{\,2} + d_i^{\,2})^2$$

$$D_i = C_i \sqrt{2K_i - \frac{B_i}{C_i}} \qquad K_i = 1 + \sqrt{1 - \frac{B_i - 1}{C_i}}$$

$$E_i = \sqrt{(C_i - \frac{B_i^{\,2}}{4})(2K_i - 4 + \frac{B_i}{C_i})} - \frac{B_i}{2} \sqrt{2K_i - \frac{B_i}{C_i}}$$

By using the above equation, we can obtain a first approximation of the group delay characteristics before executing any computation. The second term of equation (17) indicates the group delay contribution of attenuation poles at the real axis, and the fourth term means that of complex attenuation poles. These contributions are indicated as the design charts in figures 2 and 3.

In the case of actual design, first select the initial values of parameters Q_ν, Q_i, γ_ν, γ_i etc. so that the required attenuation characteristics in the stop bands may be satisfied by an estimation value, $10 \log_{10} \left| 1 + H^2 \cosh^2 \mathcal{U} \right|$ and at the same time, the required pass band delay characteristics may be satisfied by the estimation formula (7). As can be seen from the above discussion, introduction of complex attenuation poles improves, reasonably, delay distortion in the pass band and also improves attenuation in the stop bands. This procedure can be accomplished only by a simple templet method, using a design chart as indicated in figures 2 and 3.

Then, using these initial values of the parameters, compute the operating transfer factor $S(\lambda)$ with complex attenuation poles in virtue of $\phi(\lambda)$ and equation (4), and examine whether or not the requirements are actually satisfied by this $S(\lambda)$, both in the delay and in the attenuation. Although the transmission characteristics calculated from the $S(\lambda)$ with such initial values of the parameters show a good approximation, we should, in general, change slightly those values in order to improve the group delay and the attenuation characteristics. This iterative procedure can be performed by a digital computer.

After obtaining suitable $S(\lambda)$ and $\phi(\lambda)$, we must realize these characteristics by using a reactance filter with complex poles. The realization procedure will be described in the next section.

3 Synthesis of reactance filters with complex attenuation poles.

The problem considered in this section is to realize a reactance filter with complex attenuation poles as well as real frequency poles. For such a problem, many techniques have been suggested. As a simple but useful method of realization, we shall explain an application of the theory of networks using complex valued elements, or complex network theory, as an abbreviation, which was developed by one of the authors.(5)

This theory is an extension of the theory originated by R.F. Baum(8) and developed by V. Belevitch(9), in which it is permissible to use complex, rather than positive real resistors, capacitors and inductors. It is not only effective in the case of narrow band filters, as pointed out by Baum, but is strongly of use in equivalent circuit transformations, and in synthesis of two-port networks with complex transmission zeros. The detailed presentation of this theory will appear in a later publication by the same author, so for the purpose of this section we shall only introduce several useful results of this complex network theory without proofs.

In a linear, time invariant, finite, lumped, reciprocal two-port network, the cascade matrix can be written as

$$K = \frac{1}{f(\lambda)} \begin{bmatrix} a(\lambda) & b(\lambda) \\ c(\lambda) & d(\lambda) \end{bmatrix} \qquad (18)$$

where $f(\lambda)$, $a(\lambda)$, $b(\lambda)$, $c(\lambda)$ and $d(\lambda)$ are polynomials of λ, and satisfy the condition

$$a(\lambda) \cdot d(\lambda) - b(\lambda) \cdot c(\lambda) = f^2(\lambda).$$

Now, the following theorem can be established.

[Theorem 1]

If $f(\lambda) = (\lambda - \lambda_o)\, f^{(1)}(\lambda)$ and $a(\lambda_o) + b(\lambda_o) \neq 0$, then K can be factorized as follows

$$K = \frac{1}{(\lambda_o - \lambda)} \begin{bmatrix} \phi^{-1}\cdot\lambda + \lambda_o & (1+\phi)W_o\cdot\lambda \\ (1+\phi^{-1})\frac{\lambda}{W_o} & \phi\cdot\lambda + \lambda_o \end{bmatrix} K^{(1)}, \qquad (19)$$

where $K^{(1)}$ is a cascade matrix whose order is reduced by one from that of K. W_o and ϕ are expressed as follows

$$W_o = \frac{a(\lambda_o)\cdot r + b(\lambda_o)}{c(\lambda_o)\cdot r + d(\lambda_o)} \qquad r\text{: an arbitrary constant} \qquad (20)$$

$$1 + \phi = \frac{1}{\left\{\dfrac{a'(\lambda_o)\cdot r + b'(\lambda_o)}{a(\lambda_o)\cdot r + b(\lambda_o)} - \dfrac{c'(\lambda_o)\cdot r + d'(\lambda_o)}{c(\lambda_o)\cdot r + d(\lambda_o)}\right\}\cdot\lambda_o} \qquad (21)$$

By application of this theorem, we can remove the canonical section with complex valued elements from the original network as indicated in figure 4. In the figure, g, ℓ and ϕ are a complex conductance, a complex inductance and a complex transformer ratio, respectively. If λ_o is a pure imaginary value and K is a positive cascade matrix,* then the residual matrix $K^{(1)}$ is also positive cascade matrix. In this special case, theorem 1 corresponds to the Baum's procedure.

If each elements of a cascade matrix K have the properties,

$\frac{a(\lambda)}{f(\lambda)}$ and $\frac{d(\lambda)}{f(\lambda)}$ are even rational functions (with complex coefficients),

$\frac{b(\lambda)}{f(\lambda)}$ and $\frac{c(\lambda)}{f(\lambda)}$ are odd rational functions (with complex coefficients),

then K may be called a complex reactance cascade matrix. If $\frac{a(\lambda)}{f(\lambda)}$ and $\frac{c(\lambda)}{d(\lambda)}$ are positive real reactance functions, then reactance cascade matrix K becomes conventional p.r. reactance cascade matrix.

For such a complex reactance cascade matrix K, the next theorem is valid.

[Theorem 2]

For a complex reactance cascade matrix K having attenuation poles λ_o and $-\lambda_o$, if we apply theorem 1 with respect to λ_o and $-\lambda_o$ successively, then we have two canonical sections defined by the set of elements $(g_1, \phi_1\ \ell_1)$ and

---

* If a cascade matrix can be transformed to a positive (real) immittance matrix, then we shall call this cascade matrix a positive (real) cascade matrix.

(g_2, ϕ_2, ℓ_2), connected in cascade as indicated in figure 5. In this case, the following relationships exist.

$$\begin{aligned} g_1 + g_2 &= 0 \\ (1+\phi_1)+(1+\phi_2^{-1}) &= 0 \qquad (22) \\ \phi_1\ell_1 &= \phi_2^{-1}\ \ell_2 \end{aligned}$$

When K is p.r. reactance cascade matrix and λ_o is pure real or pure imaginary, removed sections can be transformed into a p.r. reactance network. This corresponds to Belevitch's cycle.

By the application of theorem 2, we can obtain 4 types of complex reactance canonical sections, corresponding to Brune's section, Darlington's E section and complex sections, as shown in figure 6.

Furthermore, when K has complex attenuation poles λ_o, $-\lambda_o$, λ^*_o, $-\lambda^*_o$ (λ^*_o means complex conjugate of λ_o), we can remove these poles by applying theorem 1, four times as illustrated in figure 7. In the figure, when the original K is p.r. reactance cascade matrix, the following relationship is valid, $\phi_1 \cdot \phi_2 \cdot \phi'_1 \cdot \phi'_2 = 1$. This corresponds to the so called Darlington's D section.

By using the preceding procedures, we are formally able to remove complex as well as real attenuation poles with ease. The resultant network consists of complex valued inductors, capacitors and resistors, which are not permissible to use in an actual filter. Therefore, we must change the element values and circuit configurations by some proper equivalent transformations. For this purpose, the following extended "Bartlett's theorem" is of great use.

[Theorem 3]

In a reciprocal two-port network N, if a Z or Y matrix exists, then this network N can always be transformed into an equivalent bridged T type network as shown in figure 8, where the ideal transformer ratio ϕ is expressed as

$$\phi = \frac{\left(\frac{Z_{22}}{Z_{12}}\right) - 1}{\left(\frac{Z_{11}}{Z_{12}}\right) - 1} \quad \text{or} \quad \frac{\left(\frac{Y_{11}}{Y_{12}}\right) + 1}{\left(\frac{Y_{22}}{Y_{12}}\right) + 1} \qquad (23)$$

As may be seen from figure 8, if the original network N can be divided into two identical subnetworks, then theorem 3 is nothing but an ordinary bisection theorem. In this sense, theorem 3 may be called an extended Bartlett's theorem.

In view of theorem 3, canonical sections indicated in figures 6 and 7 are transformed into figures 9 and 10, respectively. If the original K is a p.r. reactance cascade matrix, it can be proved that all elements in figure 9 and 10 are positive real.

In order to proceed with the synthesis of reactance filter with complex attenuation poles, we apply theorems 1 or 2 successively, along the proper order of complex or real attenuation poles. Next, transform each pair of complex sections into one bridged T section, respectively. A transformed bridged T section may have p.r. elements. The ideal transformer included in each bridged T section can be realized with real elements by a well known transformation. If realized real network containes some negative elements, we should change the order of removal of attenuation poles of this filter.

This synthesis procedure is very simple, especially for the use of a digital computer, and is effective in avoiding numerical errors.

4 Example

As an example we shall explain design procedures of a band pass filter of the 17th order. First by the templet method, suggested in the preceding section, and second by iterative computations, we determined the following parameters such that all requirements, including group delay, are satisfied,

$$k' = 0.846\ 9601\ 68$$

$$j\,q_{-3} = j\,0.716\ 9811\ 32 \qquad \gamma_{-3} = 1$$

$$j\,q_{-2} = j\,0.798\ 7421\ 38 \qquad \gamma_{-2} = 1$$

$$j\,q_{-1} = j\,0.821\ 8029\ 35 \qquad \gamma_{-1} = 1$$

$$j\,q_{1} = j\,1.015\ 0943\ 39 \qquad \gamma_{1} = 1$$

$$j\,q_{2} = j\,1.029\ 7 \qquad \gamma_{2} = 1$$

$$d_\beta + j\,q_\beta = 0.062 + j\,0.925 \qquad \gamma_\beta = 1$$

$$-d_\beta + j\,q_\beta = -0.062 + j\,0.925 \qquad \gamma_\beta^{*} = 1$$

$$q_\alpha = 0.3 \qquad \gamma_\alpha = -\frac{1}{2}$$

$$\gamma_o = \frac{1}{2}$$

$$\gamma_\infty = 1$$

A_e = 18.0 db (Minimum echo loss in the pass band)

In the above table, q_α with γ_α is merely a temporary pole, which is necessary for the design of a band pass filter of odd order,$^{(4)}$ and should be removed after accomplishing approximations. q_β and d_β in the above list are complex attenuation poles which improve both group delay in the pass band and attenuation in the stop bands.

As already stated by some of the authors of this paper,$^{(7)}$ in order to avoid accumulations of computational errors, all computations should be done on the λ-plane by the following transformation

$$p^2 = \alpha(\lambda^2 + Q_o^{\ 2}),$$

$$\alpha = 1 - k'^2 \quad ,$$

$$Q_o = -\frac{k'^2}{1 - k'^2} \quad .$$

Then, we have the operating transfer factor $S(\lambda) = H \cdot \frac{(\lambda)}{f(\lambda)}$ as follows:

$$\mathscr{g}(\lambda) = \left(\sqrt{\lambda^2 + Q_o^2} + C_o\right) \prod_{\nu=1}^{8} \left(\lambda^2 + b_\nu \cdot \sqrt{\lambda^2 + Q_o^2} + C_\nu\right),$$

$$f(\lambda) = \sqrt{\lambda^2 + Q_o^2} \cdot \left(\lambda^2 + Q_\beta^2\right) \cdot \left(\lambda^2 + Q_\beta^{*2}\right) \cdot \prod_{\substack{\nu=-3 \\ \nu \neq 0}}^{2} \left(\lambda^2 + Q_\nu^2\right),$$

where

$$Q_i^2 = \frac{q_i^2 - k'^2}{1 - k'^2} \qquad i = -3.\ -2.\ -1.\ 1.\ 2.$$

$$Q_\beta^2 = \frac{-(d_\beta + j\ \beta)^2 - k'^2}{1 - k'^2} \quad .$$

The values of Q_i's, b_ν's and C_ν's are listed below:

$Q_0^2 = -2.537\ 838383$

$Q_{-3}^2 = -0.148\ 523641$

$Q_{-2}^2 = -0.280\ 736411$

$Q_{-1}^2 = -0.719\ 170321$

$Q_1^2 = 1.107\ \ 608661$

$Q_2^2 = 1.213\ \ 268327$

$Q_\beta^2 = 0.475\ 6251\ 32 + j\ 0.405\ 790063$

$C_0 = 5.642\ 7376\ 28 \times 10^{-1}$

$b_1 = 1.451\ 6664\ 6 \times 10^{-2}$ $\qquad C_1 = -1.476\ 2176\ 5 \times 10^{-2}$

$b_2 = 5.008\ 3068\ 6 \times 10^{-2}$ $\qquad C_2 = 3.845\ 8403\ 3 \times 10^{-2}$

$b_3 = 8.985\ 2771\ 6 \times 10^{-2}$ $\qquad C_3 = 1.758\ 9099\ 98 \times 10^{-1}$

$b_4 = 1.023\ 3775\ 43 \times 10^{-1}$ $\qquad C_4 = 3.788\ 9422\ 27 \times 10^{-1}$

$b_5 = 1.030\ 3607\ 36 \times 10^{-1}$ $\qquad C_5 = 5.808\ 0078\ 00 \times 10^{-1}$

$b_6 = 4.864\ 3638\ 71 \times 10^{-2}$ $\qquad C_6 = 9.647\ 4938\ 83 \times 10^{-1}$

$b_7 = 9.445\ 9084\ 05 \times 10^{-2}$ $\qquad C_7 = 8.046\ 3939\ 99 \times 10^{-1}$

$b_8 = 1.240\ 5201\ 95 \times 10^{-2}$ $\qquad C_8 = 1.016\ 0855\ 32$

The attenuation and group delay characteristics calculated from this $S(\lambda)$ are indicated in figures 10 and 11, respectively. In order to estimate the contribution of the complex attenuation poles $\pm d_\beta \pm jq_\beta$, the attenuation characteristics of a Tchebycheff filter having the same parameters as the designed one, except complex poles, are indicated by dotted lines in figure 10, and the group delay characteristics of the Techebycheff filter which has the same attenuation as the designed one, but has no zeros and poles on the right half of the p-plane.

By observing these figures, we can estimate the amount of the improvements due to the complex attenuation poles, but only for the group delay, by more than 130 μs in the main part of the pass band, but also for the attenuation by more than 10 dB in almost all of the stop bands. This result is somewhat surprising in that the improvements are fairly good both for delay and attenuation as one would expect in the comparison with the case of the ordinary all pass delay equalizer.

Next, the circuit configuration with complex valued elements is synthesized by the procedures described in section 3. All computations of synthesis procedures are simply made by a digital computer only using a complex interpretive subroutine. Complex valued elements obtained here are transformed into positive real values by applying the transformation described in theorem 3. The calculated results of this procedure are shown in figure 12. The ideal transformer included in the network can be replaced by three actual positive capacitors as shown in figure 13.

5 Conclusion

A useful design method for the filter, which exhibits Tchebycheff performance and at the same time has a fairly flat group delay in the pass band, is presented. This design is performed essentially by introductions of complex attenuation poles into Tchebycheff filters. In the approximation stage, a simple and useful design chart method is described. Although approximations should be accomplished by iterative computation, using a digital computer, such a templet method is still powerful enough to select the initial valuesof the parameters. In the realization stage, several interesting results of network theory with complex valued elements are given. Lastly, an extended Bartlett's theorem is reported. By using these results, synthesis of reactance filters, with complex attenuation poles, can easily and simply be performed, provided that a complex interpretive subroutine is prepared. The design example shows us that the method suggested in this paper is most effective for practical purposes. It is hoped that the present work will provide some useful technique to the field of transmission filter design.

6 Acknowledgement

The authors wish to acknowledge the direction of Mr. Haruo Tomonori, Manager of the Component Engineering Division, Mr. Masanori Kogo and others of the Nippon Electric Co. in preparing this paper. Their sincere gratitude is also extended to all members of the Network Design Group of our company for their valuable assistance.

References

(1) For instance, see Louis Weinberg "Network Analysis and Synthesis" McGraw Hill 1962.

(2) J.E. Colin "Modelage du temps de propagation de groupe d'un filtre à caracteristiques données d'affaiblissment composite" Câbles et Transmission No. 3. 1963 pp. 165-173.

(3) H. Watanabe et al "Group delay characteristics of Tchebycheff filters" National Convention of I.E.C.E. Japan No. 7. Nov. 1960 (in Japanese).

(4) H. Watanabe "Approximation Theory for Filter-Networks" IRE Trans. on Circuit Theory Vol. CT-9. No. 3. pp. 341-356. Sept. 1961.

(5) H. Watanabe "A consideration on the theory of cascade synthesis of two-port networks" (in Japanese) PTG on Circuit Theory Japan. Jan. 19, 1963.

(6) H. Watanabe et al "Design of filters with complex attenuation poles" (in Japanese) Joint convention of Electrical Engineering No. 7 April 1963.

(7) T. Iedokoro, T. Tsuchiya and H. Watanabe "A new calculation method for the design of filters by digital computer with the special consideration of the accuracy problem" IEEE International Convention Record. March 1963.

(8) R.F. Baum "A Modification of Brune's Method for Narrow-Band Filters" IRE Trans. on Circuit Theory. Vol. CT-5 pp. 264-267. Dec. 1958.

(9) V. Belevitch "An Alternative Derivation of Brune's Cycle" IRE Trans. on Circuit Theory. Vol. CT-6 pp. 389. Dec. 1959.

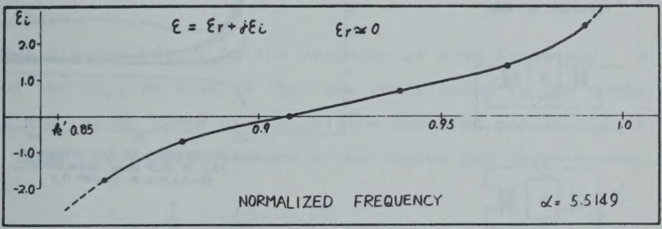


Fig. 1 AN EXAMPLE OF THE MAGNITUDE OF ε

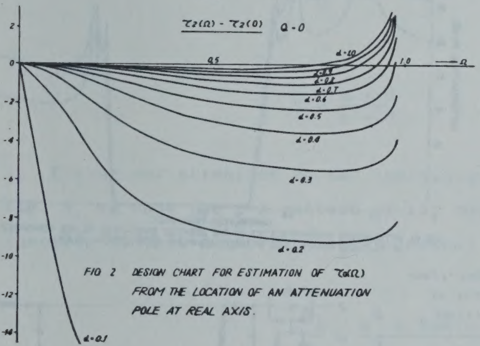


FIG 2 DESIGN CHART FOR ESTIMATION OF τ2(Ω) FROM THE LOCATION OF AN ATTENUATION POLE AT REAL AXIS.

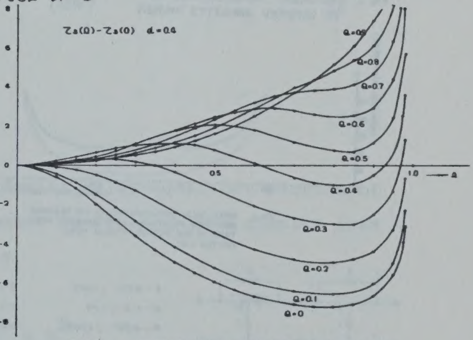
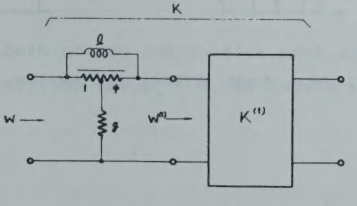


FIG 3 DESIGN CHART FOR ESTIMATION OF τ2(Ω) FROM THE LOCATION OF AN ATTENUATION POLE IN COMPLEX DOMAIN.



$$l = (1+\phi)\frac{W_0}{\lambda_0}$$

$$g = (1+\phi^{-1})\frac{1}{W_0}$$

$$1+\phi = \frac{W_0}{\lambda_0 \cdot W_0'}$$

$$W_0 = W(\lambda_0)$$

FIG 4 AN ILLUSTRATION OF THE THEOREM 1

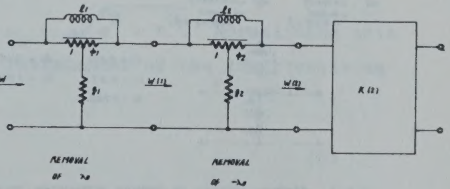
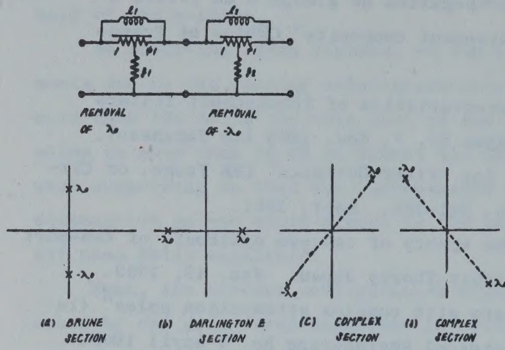


FIG 5 AN ILLUSTRATION OF THE THEOREM 2


REMOVAL OF λ0
REMOVAL OF -λ0
λ0
-λ0

(a) BRUNE SECTION (b) DARLINGTON E SECTION (C) COMPLEX SECTION (d) COMPLEX SECTION

FIG 6 REMOVAL OF ATTENUATION POLES

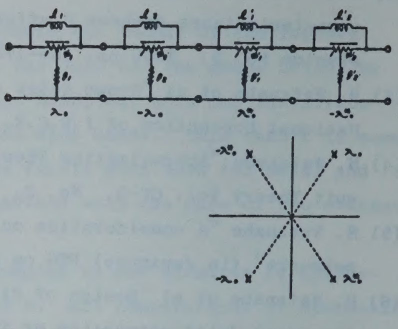


FIG 7 REMOVAL OF CONJUGATE ATTENUATION POLES.

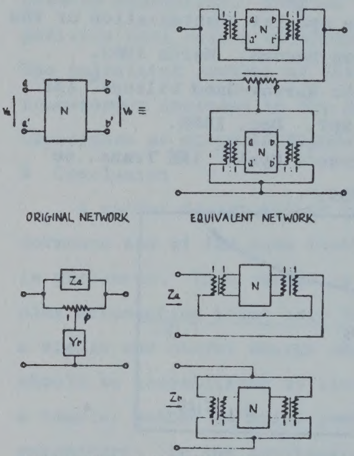

N
ORIGINAL NETWORK
EQUIVALENT NETWORK
Za
Yr
Zr

FIG 8 EQUIVALENT TRANSFORMATION BY THE EXTENDED BARTLETT'S THEOREM

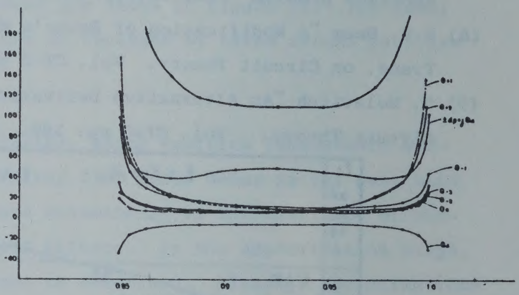


FIG 9 CONTRIBUTION OF THE EACH ATTENUATION POLE TO GROUP DELAY CHARACTERISTICS. (CALCULATED BY EQUATION 17)

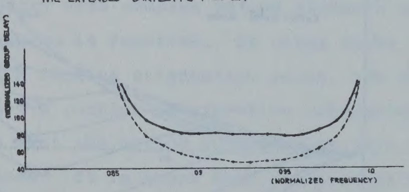


FIG 11 GROUP DELAY CHARACTERISTICS OF THE DESIGNED BAND-PASS FILTER (SOLID LINE) COMPARED WITH THOSE IN THE CASE OF MINUMUM PHASE (DOTTED LINE)

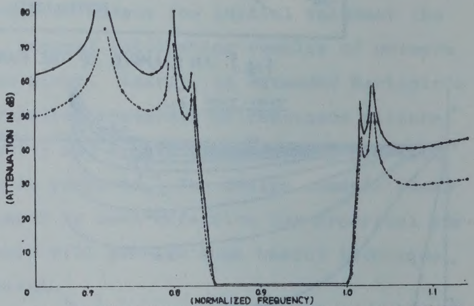


FIG 10 ATTENUATION CHARACTERISTICS OF THE DESIGNED BAND-PASS FILTER COMPARED WITH THOSE OF A FILTER WITHOUT COMPLEX ATTENUATION POLES.

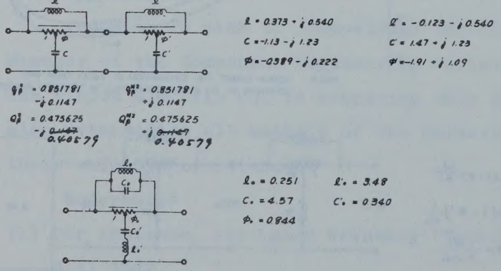

ℓ = 0.373 + j 0.540
C = -1.13 - j 1.23
φ = -0.389 - j 0.222
ℓ' = -0.123 - j 0.540
C' = 1.47 + j 1.23
φ' = -1.91 + j 1.09
Q = 0.851781 - j 0.1147
Q = 0.851781 + j 0.1147
Q = 0.475625 - j ~~0.1147~~ 0.40579
Q = 0.475625 + j ~~0.1147~~ 0.40579
ℓ0 = 0.251
C0 = 4.57
φ0 = 0.844
ℓ'0 = 3.48
C'0 = 0.340

FIG 12 REMOVAL PROCESS OF COMPLEX ATTENUATION POLES AND A TRANSFORMATION TO PRACTICAL NETWORK

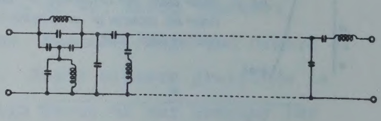


FIG 13 FINAL NETWORK CONFIGURATION

DELAY AND ATTENUATION EQUALIZERS WITH LOSSY REACTANCES

PHILIP R. GEFFE
Axel Electronics, Inc.
Jamaica, New York

Abstract Incidental dissipation in all-pass lattices is corrected with the Dagnall-Rounds equivalence. Unbalanced all-pass structures (or lattices) may be predistorted, or compensated by a combination of pre- and post-correction.

Attenuation equalizers are predistorted in constant-resistance networks to give the required response with a flat loss. Design details are discussed.

Attenuation of Lossy All-Pass Lattices In fig. 1 we see a bilinear all-pass lattice network with lossless reactances. Its complex loss function is given by (1).

$$\frac{e_1}{e_2} = \frac{\left(\frac{1}{a}\right) S + 1}{\left(-\frac{1}{a}\right) S + 1} \tag{1}$$

If losses are introduced uniformly, the p-z pattern of the response is displaced a distance, d, to the left, as shown in fig. 2, with the lossy lattice section also shown there.

Clearly a dip now occurs in the response at zero frequency. To evaluate the magnitude of the dip, we observe that the resistances in the network have reciprocal values, as do those of the lattice matching pad in fig. 3. At zero frequency, then, the lattice becomes a resistance pad whose insertion loss is given by (2)

$$\alpha_o = 20 \log K \qquad [\text{db}] \tag{2}$$

where

$$K = \frac{a + d}{a - d} \tag{3}$$

Fixing our attention on the lossy biquadratic delay-equalizer section of fig. 4, we show the p-z pattern of its response in fig. 5. The complex loss function of this network (without losses) is

$$\frac{e_1}{e_2} = \frac{s^2 + 2aS + a^2 + b^2}{s^2 - 2aS + a^2 + b^2} \tag{4}$$

Both series and lattice arms are resonant at $\omega_o = \sqrt{a^2 + b^2}$. Normalizing this section for $\omega_o = 1$, we obtain the equivalent resistances of the lossy coils as

$$\left.\begin{aligned} r &= \frac{d}{2a} \\ R &= \frac{2a}{d} \end{aligned}\right\} \tag{5}$$

Since we again have reciprocal resistances in series and lattice arms, the insertion loss at ω_o can be calculated from fig. 3 and eq. (2) giving

$$K = \frac{2a + d}{2a - d} \tag{6}$$

It is easy to show[3] that this is the attenuation peak of the section, whose amplitude response is shown in fig. 6. The nonflat amplitude response caused by incidental dissipation is not the only reason for concern, since the phase and delay are also distorted. It follows that some method of compensation must be devised.

The Dagnall-Rounds Equivalence* As given by Dagnall and Rounds[1], the lossy lattice of fig. 7a has the same transfer function, within a constant multiplier, as the lattice of fig. 7b. This equivalence depends only on Z_a and Z_b being reciprocal, realizable, RIC impedances, and on the resistors, R_1 and R_2, being reciprocals also (Observe that the network is normalized for 1 ohm terminations).

Thus we find that Z'_a, the series branch of fig. 7a, is

$$Z'_a = R_2 \left(\frac{Z_a + R_1}{Z_a + R_2} \right) \tag{7}$$

Since R_1 and R_2 are reciprocals,

$$Z'_a = \frac{Z_a + R_1}{R_1 Z_a + R_1} \tag{8}$$

Finally, the transfer impedance, Z'_{21}, of the lossy lattice, is found to be

$$\begin{aligned} Z'_{21} &= \frac{1 - Z'_a}{1 + Z'_a} \\ &= \left(\frac{1 - R_1}{1 + R_1}\right)\left(\frac{1 - Z_a}{1 + Z_a}\right) \\ &= \left(\frac{1 - R_1}{1 + R_1}\right) Z_{21} \end{aligned} \tag{9}$$

where Z_{21} is the transfer impedance of the lattice in fig. 7b. Accordingly, the flat loss of the lossy lattice is given by (2), taking K as

$$K = \frac{1 + R_1}{1 - R_1} \tag{10}$$

Since Z'_a and Z'_b are also reciprocal impedances, the constant-resistance property of the lattice is preserved.

---

* This equivalence is stated, rather badly, in reference 1. Our purpose here is to give design details.

When it is desired to apply this lattice equivalence to the network of fig. 4, and if the element-Q is sufficiently high, then we find that $R_2 \gg R_1$. Fig. 7a then reduces to fig. 8, where R_1 and R_2 are physical resistors, but r and R simply represent coil losses. At ω_o, the lattice essentially reduces to the matching pad of fig. 3, with R in the series arm, and r in the lattice arm. At $\omega = 0$, the same attenuation is given by the added resistances. Since the DC resistance of the series coil affects the attenuation at zero frequency, this value should be subtracted from R_1. To preserve flatness at hich frequencies, it is necessary to provide the capacitor of the series branch with the same series resistance as its associated coil.

$$\left.\begin{aligned} R_1 &= r - R_{dc} \\ R_2 &= R \end{aligned}\right\} \qquad (11)$$

The high-Q case therefore permits compensation of losses merely by the addition of six resistors to the otherwise unmodified lattice.[**]

From a practical point of view, these applications of the Dagnall-Rounds equivalence are only approximate. In some cases, they will fail to give satisfaction because they neglect winding (series) resistance in the series coil, and core loss (parallel) resistance in the lattice coil. When high accuracy is desired, this method often necessitates increasing coil dissipation with physical resistors, so that the neglected losses become a small fraction of the total losses.

An ideal transformer must be used when these networks are applied to unbalanced circuitry. This is shown in fig. 9. The difficulty in obtaining close-coupled coils becomes severe at carrier frequencies, when the methods outlined below may be found more useful.

Predistortion of All-Pass Networks   Precorrection for effects of lossy reactances in all-pass networks is not so straightforward as in ladder networks, but is accomplished with the aid of attenuation equalizers.[***]

Suppose that the design objective is to place a response zero at a distance, +a, from the jω-axis, and a response pole at the same distance in the other (left) half-plane. A lossless design is first obtained with its singularities at a distance (a + d) on either side, as shown in fig. 10. Thus, we plan on p_1 and z_1 as the final pole and zero, but initially design for p_2 and z_2.

[**] The high-Q case is mentioned at the end of ref. 2.

[***] This method may not be new. Reference 1 states that a lossy delay equalizer "of modified constants" can be compensated with an attenuation equalizer; but no design clues are given. It is, perhaps, possible that the authors had in mind one of the methods described here.

The next step is to obtain an attenuation equalizer having a pole at p_3, and a zero at p_2. The result of both of these design steps is a network which has a zero at z_2 and a pole at p_3, when built with lossless reactances (Of course, the pole and zero at p_2 cancel out). This p-z pattern is now a distance, d, to the right of the location which is desired in the final result. It is now only necessary to add uniform dissipation to all the reactances of both networks, in the amount d, and the p-z pair will slide into its desired location.

This method is equally effective in both bilinear and biquadratic sections, and in both balanced and unbalanced circuitry. For accurate results, it is essential that the added dissipation be substantially uniform. This requirement may sometimes occasion odd practices, such as wiring physical resistors in series and parallel with mica capacitors, as dictated by circumstances.

Combined Precorrection and Postcorrection A very similar compensation scheme is conceptually different. A lossless delay equalizer is first designed, exactly as in the first step above. Adding dissipation to this network causes it to have a zero at Z_1 and a pole at p_4. This is cascaded with a practical attenuation equalizer having an actual pole at p_1 and a zero at p_4. This yields the prescribed transfer function in a physical network which may be different from the result of the previous design procedure.

In practical problems, we can build a cascade of a number of delay equalizer sections. Taking care that the dissipation is uniform, and that the right-half-plane zeros are correctly located, the resulting irregular amplitude response is flattened out with any attenuation equalizer which will serve. So long as this process (which may be highly empirical) leads to an all-pass response, then it may safely be concluded that the response poles have been accurately compensated. The validity of this procedure rests on the fact that the p-z pattern of a minimum-phase network can be determined solely from its amplitude response alone.

Attenuation Equalizers The design of attenuation equalizers for accurate service in the above applications requires that some scheme of compensation for lossy reactances should be devised for them also. Other applications give rise to the same need. For example, band-pass filters built with inadequate element-Q suffer a severe rounding-off of the response at the pass band edge. Postcorrection of this phenomenon, with attenuation equalizers, also encounters element-Q problems - especially in narrow-band filters. A little practical experience will confirm this observation on a note of urgency.

The design of conventional equalizers can be recast into a more convenient form by synthesizing them directly from the p-z pattern of fig. 11. The complex loss function is

$$A(s) = \frac{\left(\frac{1}{a}\right)S + 1}{\left(\frac{1}{b}\right)S + 1} \qquad (12)$$

With 0 db at $\omega = 0$, the loss at high frequencies (total equalization) is

$$\left.\begin{aligned} \alpha_o &= 20 \log K \qquad [\text{db}] \\ K &= \frac{b}{a} \end{aligned}\right\} \qquad (13)$$

and the square of the insertion ratio at real frequencies is

$$\left|A(j\omega)\right|^2 = K^2 \left(\frac{\omega^2 + a^2}{\omega^2 + a^2 K^2} \right) \qquad (14)$$

Obviously, we get an attenuation (db) of $\frac{1}{2}\,\alpha_o$ at a frequency, ω_o, such that

$$K^2 \left(\frac{\omega_o^2 + a^2}{\omega_o^2 + a^2 K^2} \right) = K \qquad \text{or when}$$

$$a^2 K - \omega_o^2 = 0 \qquad (15)$$

Desiring to normalize at $\omega_o = 1$, we have

$$\left.\begin{aligned} a &= \frac{1}{\sqrt{K}} \\ b &= \sqrt{K} \end{aligned}\right\} \qquad (16)$$

This synthesis leads to the network, and response, as shown in fig. 12. The attenuation, $\alpha(\omega)$, at a frequency, ω, is easy to calculate from $B(\omega)$:

$$B(\omega) = \frac{\omega^2 + K}{\omega^2 + \frac{1}{K}} \qquad (17)$$

$$\alpha(\omega) = \alpha_o - 10 \log B(\omega) \qquad [\text{db}] \qquad (18)$$

$$\alpha(\tfrac{1}{\omega}) = 10 \log B(\omega) \qquad [\text{db}] \qquad (19)$$

This is a consequence of the fact that

$$\alpha(\omega) + \alpha(\tfrac{1}{\omega}) = \alpha_o \qquad [\text{db}] \qquad (20)$$

which follows identically from the above relations.

It is often convenient to replace the resistance tee pad of fig. 12 with the equivalent bridged-tee pad, as shown in fig. 13. This is so because the bridged-tee pad can be used to absorb parasitic losses in both high-pass and band-attenuation sections, as shown in fig. 14. It is apparent, however, that this resource fails to accommodate the low-pass and band-pass sections. For these cases, we have developed a different method.

Predistorted Attenuation Equalizers If a low-pass attenuation equalizer (unnormalized) has the p-z pattern of fig. 11, we can calculate the insertion ratio, K, and the half-attenuation frequency, ω_o:

$$\left.\begin{aligned} K &= \frac{b}{a} \\ \omega_o &= \sqrt{ab} \end{aligned}\right\} \tag{21}$$

In order to predistort this network, we first displace the p-z pattern a distance, d, to the right. This is realized in an equalizer which has lossless reactances. When uniform losses are added to this network, the originally-desired result is achieved.

The lossless, predistorted, design therefore has modified values of k and ω_o:

$$\left.\begin{aligned} K' &= \frac{b-d}{a-d} \\ \omega_o' &= \sqrt{(a-d)\ (b-d)} \end{aligned}\right\} \tag{22}$$

and this leads to a realizable network if and only if

$$d < a \tag{23}$$

These last two relations are sufficient for design purposes, but a more explicit set of design formulas go with the network of fig. 15. The element values are

$$L = C = \frac{\sqrt{K} - \frac{1}{\sqrt{K}}}{\left(\sqrt{K} - d\right)\left(\frac{1}{\sqrt{K}} - d\right)} \tag{24}$$

$$R_1 = \frac{\sqrt{K} - \frac{1}{\sqrt{K}}}{\sqrt{K} + {}^{1}/\sqrt{K} - 2d} \tag{25}$$

$$R_2 = \frac{2(\sqrt{K} - d)({}^{1}/\sqrt{K} - d)}{(\sqrt{K} - d)^2 - (\frac{1}{\sqrt{K}} - d)^2} \tag{26}$$

$$R_B = \frac{\sqrt{K} - \frac{1}{\sqrt{K}}}{\frac{1}{\sqrt{K}} - d} \qquad (27)$$

The bridging arm and the shunt arm are reciprocal impedances, so that the modified sections still have the constant resistance property at both ends. They have the response given in fig. 12, except for a flat loss, α_F:

$$\alpha_F = 20 \log \frac{\sqrt{K} - d}{\sqrt{K} - dK} \qquad [\text{db}] \qquad (28)$$

References

1. C. H. Dagnall & P. W. Rounds, "Delay Equalization of Eight-Kilocycle Carrier Program Circuits," B.S.T.J., vol. 28, April 1949, pp. 181-195.
2. F. S. Farkas, F. J. Hallenbeck, & F. E. Stehlik, "Band Pass Filter, Band Elimination Filter and Phase Simulating Network for Carrier Program Systems," B.S.T.J., vol. 28, April 1949, pp. 196-220.
3. R. Saal & K. Antreich, "Zur Realisierung von Reaktanz-Allpass-Schaltungen," Frequenz, Band 16, Nr. 12, Dec. 1962, pp. 469-477, andBand 17, Nr. 1, Jan. 1963, pp. 14-22.

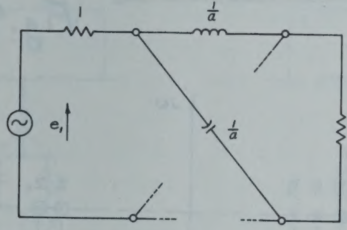

1
$\frac{1}{a}$
$\frac{1}{a}$
e_1
e_2
1

Fig. 1

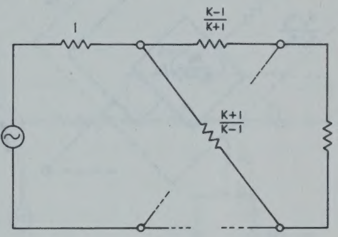

1
$\frac{K-1}{K+1}$
$\frac{K+1}{K-1}$
1

Fig 3

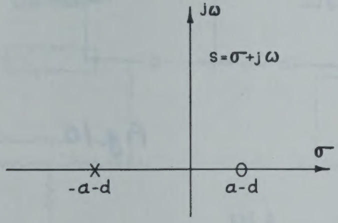
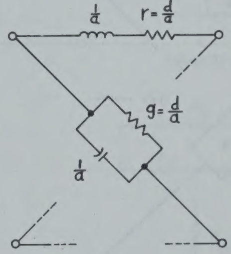

jω
s = σ + jω
σ
-a-d
a-d

Fig 2


$\frac{1}{a}$
$r = \frac{d}{a}$
$g = \frac{d}{a}$
$\frac{1}{a}$

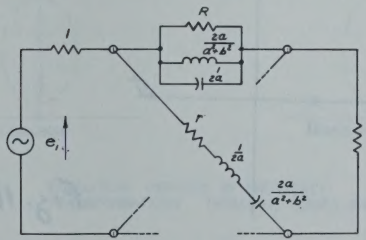

1
R
$\frac{2a}{a^2+b^2}$
$\frac{1}{2a}$
r
$\frac{1}{2a}$
$\frac{2a}{a^2+b^2}$
e_1
e_2
1

Fig 4

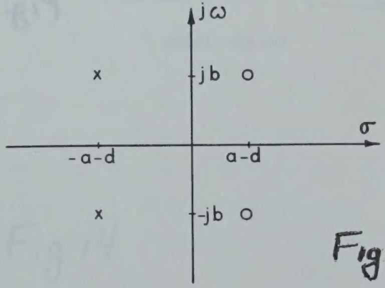

jω
x
jb
o
σ
-a-d
a-d
x
-jb
o

Fig. 5

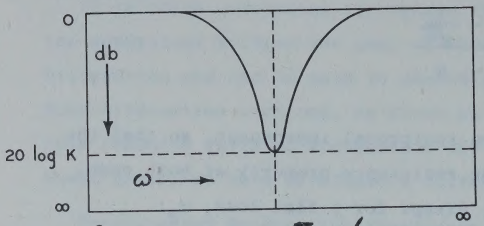


Fig 6

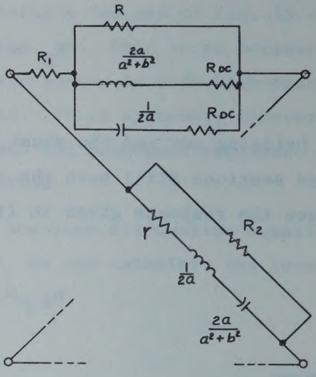

R
R1
2a/(a²+b²)
RDC
1/2a
RDC
r
R2
1/2a
2a/(a²+b²)

Fig. 8

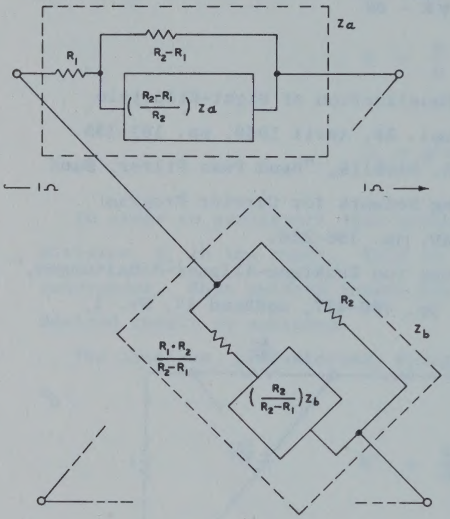

R1
R2-R1
Za
((R2-R1)/R2) Za
1Ω
1Ω
R2
Zb
(R1·R2)/(R2-R1)
(R2/(R2-R1)) Zb

Fig 7a

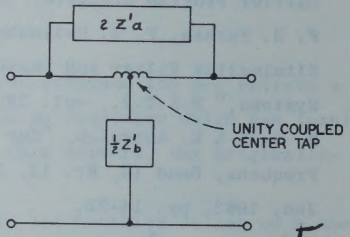

2 Z'a
½ Z'b
UNITY COUPLED
CENTER TAP

Fig. 9

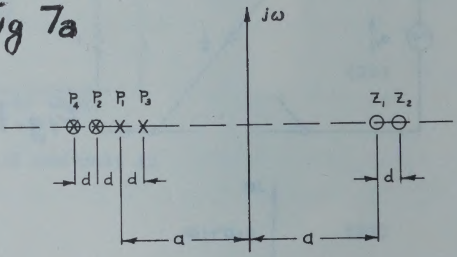

jω
P4
P2
P1
P3
Z1
Z2
d
d
d
d
a
a

Fig. 10

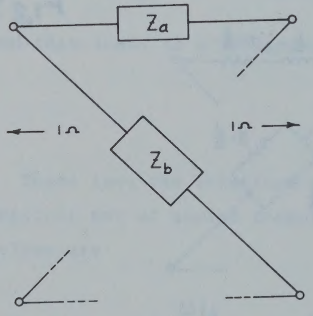


Fig. 7b

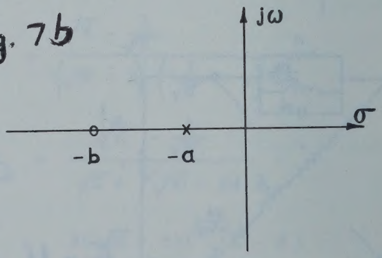
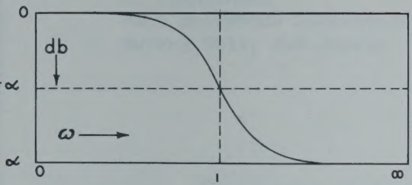


Fig. 11



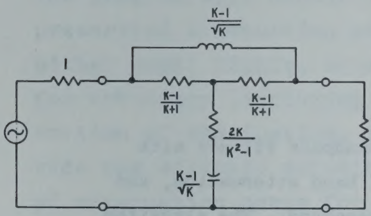

1
K-1/√K
K-1/K+1
K-1/K+1
2K/K²-1
K-1/√K
1

Fig. 12

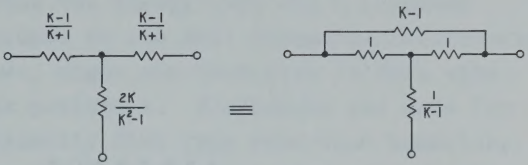
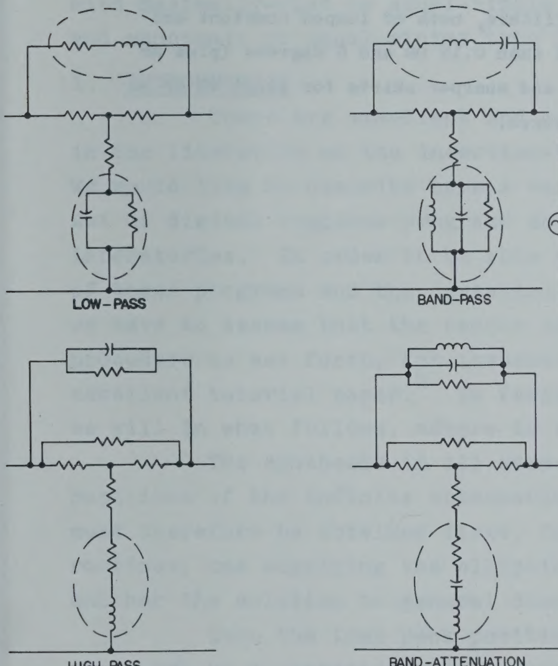

K-1/K+1
K-1/K+1
2K/K²-1
≡
K-1
1
1
1/K-1

Fig. 13


LOW-PASS
BAND-PASS
HIGH-PASS
BAND-ATTENUATION

EQUALIZERS OBTAINED BY FREQUENCY TRANSFORMATIONS (PARASITIC LOSSES SHOWN)

Fig. 14

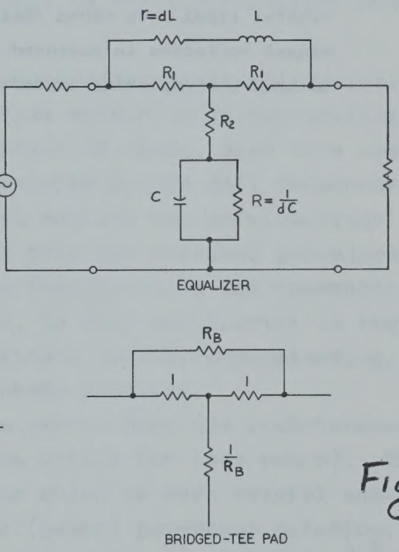

r=dL
L
R₁
R₁
R₂
C
R = 1/dC
EQUALIZER
R_B
1
1
1/R_B
BRIDGED-TEE PAD

Fig 15

LINEAR-PHASE BAND-PASS FILTER

R. M. LERNER
M.I.T. Lincoln Laboratory
Lexington 73 Massachusetts

ABSTRACT

An algorithm is given for design of bandpass filters with uniform passband transmission, high out of band attenuation, and negligible - phase distortion across the passband. The algorithm gives component values directly and is equally applicable to lumped constant and crystal designs losses and strays are easily taken into account. Experimental filters, both of lumped constant and crystal types have shown less than 0.15 DB and 5 degrees (plus or minus) variation in passband and sharper skirts for first 40 or 50 DB than corresponding butterworth.

THEORETICAL BASIS OF A DIGITAL COMPUTER PROGRAM PACKAGE FOR FILTER SYNTHESIS

G. SZENTIRMAI
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

ABSTRACT

The paper describes the theory used for a computer program package being developed at the Bell Telephone Laboratories. The program will handle low-, high- and band-pass filters with prescribed attenuation peak positions. Provisions are made for either equal ripple, or maximally flat type pass-band behavior, for arbitrary (including extreme) termination ratio and predistortion of dissipation, if required. Additional subroutines provide the elliptic solution, or the required number and positions of attenuation peaks for prescribed arbitrary suppression requirements.

A modified program provides low- and band-pass filters with maximally-flat or equal-ripple type delay in their pass-band and monotonic or equal-minima type loss in their suppression band.

1. INTRODUCTION

There are excellent and complete descriptions available in the literature on the insertion-loss method of filter synthesis.[1,2] We would like to describe here a variant of these, used in a new set of digital computer programs developed in the Bell Telephone Laboratories. In order to be able to explain the peculiarities of these programs and the deviations from the accepted procedures, we have to assume that the reader is familiar with the conventional procedure as set forth, for instance, by Saal and Ulbrich in their excellent tutorial paper.[2] To facilitate further understanding, we will in what follows, adhere to their notation.

The synthesis in all cases starts from the predetermined positions of the infinite attenuation points (or loss peaks). These must therefore be obtained first, for which we have several subroutines, one supplying the elliptic (Cauer) parameter solution, another the solution to general discrimination requirements.[3,4]

Once the loss peak positions are given, the first step is to set up a characteristic function embodying these peaks and exhibiting equal-ripple (or maximally-flat) type behavior in the pass band. This characteristic function can be expressed very simply in terms of a transformed variable, φ, where:

$$\varphi^2 = \lambda^2/(1+\lambda^2) \qquad \lambda = j\omega/\omega c \qquad (1a)$$

in the case of a low-pass filter, or

$$\varphi^2 = (\lambda^2+\beta^2)/(1+\lambda^2\beta^2) \qquad \lambda = j\omega/\sqrt{\omega_c\omega_{-c}} \quad \beta^2 = \omega_c/\omega_{-c} \qquad (1b)$$

in the case of a band-pass filter. Here ω_c and ω_{-c} are the end(s) of the pass band of the filter.

In terms of this variable, the characteristic functions of low- and band-pass filters are nearly identical in form. This variable has been used by many designers to set up the characteristic function and some have used it up to and including the step of determining the natural modes of the network.[5] From the mapping properties of the transformations (1a)-(1b) it is clear why this variable is superior for this purpose to the conventional frequency variable. The pass bands in both the low- and band-pass cases are transformed onto the whole imaginary axis in the transformed φ-plane. The natural modes which are especially close to each other near the edges of the pass band, are widely separated in the φ-plane, and consequently their extraction represents a considerably simpler problem. Furthermore a certain number of accurate significant digits on the φ-plane usually represents more accurate digits in the λ-plane. This indicates the desirability of retaining this variable even after the natural modes have been determined. In the latter stages of the synthesis, the computation in the λ variable is known to be very sensitive to numerical accuracy, and the accumulation of error can quickly lead to meaningless numbers. However, because of the greater accuracy the digits represent in the transformed variable, we can expect a synthesis in terms of this variable to be less prone to loss of accuracy through the computations.

Previously, this variable has also been used for the actual synthesis step* but it assumed the impedance functions of the network to be known.[6] To combine the two ends of the synthesis, we only needed to find the bridging step of obtaining the impedance functions from the natural modes in terms of the transformed variable φ without transforming them back into the λ plane first.

---

*To be sure, the filter was of the image parameter type, but this does not make any difference as far as the actual synthesis is concerned. It only made the step of obtaining the impedance functions much simpler.

This done, we only had to perform the analogous steps in the band pass case to have a synthesis method superior in every respect to methods known previously, at least as far as numerical accuracy and adaptability to digital computer programs are concerned. The inevitable price to be paid was the fact that the low- and band-pass filter programs became so different as to necessitate their complete divorce. This was also required, however, in order to keep the input to the programs from becoming excessively complex.

2. LOW-PASS FILTER

Let us now consider the low-pass filter program in some detail. Once the characteristic function is known, the polynomial containing the natural modes and their negatives, can be obtained exactly as in the λ plane. Factoring this (even) polynomial into quartic factors of the form

$$\left(\varphi^4 + a_i\varphi^2 + b_i\right) \tag{2}$$

The natural modes are separated by determining the expression

$$\left(A_i\varphi^2 + B_i\varphi\sqrt{1-\varphi^2} + C_i\right) \tag{3}$$

where the constants in (3) are given in terms of the constants in (2):

$$A_i = \sqrt{1+a_i+b_i} - \sqrt{b_i}$$

$$B_i = \sqrt{2\sqrt{b_i(1+a_i+b_i)} - (a_i+2b_i)} \tag{3a}$$

$$C_i = \sqrt{b_i}$$

Multiplying the factors of the form (3) back together again we obtain a function of the form

$$E(\varphi) = \sqrt{1-\varphi^2}\, E_1(\varphi) + E_2(\varphi) \tag{4}$$

where $E_1(\varphi)$ and $E_2(\varphi)$ are polynomials in φ, one even, the other odd, corresponding to the even and odd parts respectively of the polynomial containing the natural modes in the λ plane. These two polynomials, together with the characteristic function suffice to obtain the impedance or admittance functions of the network.*

---

*Depending upon the prescribed value of the termination ratio, modifications of some of these polynomials might be necessary, which are all performed automatically by the computer. In case the modifications result in complex zeros of the reflection coefficient, these are selected such that the sum of their real parts is as close to zero as possible. Extreme termination ratio (open- or short-circuit on one side) is also available.

We will not give the expressions but they have a close resemblance to their counterparts in terms of the variable λ. Finally the actual synthesis is an adapted version of the process known as zero shifting. Although the derivation of the expressions for this process is also available in the literature,[6] let us see just one step to show the essential features of the synthesis. Assume that Z_1 is the input (open- or short-circuited) impedance of the filter, in which case Z_1/λ is an even rational function of λ and hence it is also an even rational function of φ:

$$Z_1/\lambda = w_1(\varphi) \tag{5}$$

Assume also that the configuration required is that of Fig. 1, where L_2 and C_2 resonate at ω_2 which is prescribed.

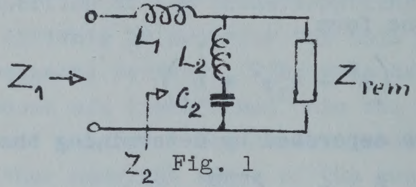


Fig. 1

At $\lambda^2 = -\omega_2^2$ or $\varphi^2 = \frac{\omega_2^2}{\omega_2^2 - 1} = \varphi_2^2$ Z_1/λ will be equal to L_1, hence:

$$L_1 = Z_1/\lambda \Big|_{\varphi=\varphi_2} = w_1(\varphi_2) \tag{6}$$

and consequently

$$Z_1/\lambda - L_1 = w_1(\varphi) - w_1(\varphi_2) = (\varphi^2 - \varphi_2^2)w_2(\varphi) = Z_2/\lambda \tag{7}$$

since $w_1(\varphi)$ is an even function. The resonant circuit has an admittance using Eq. (1a):

$$\frac{C_2\lambda}{1 + L_2C_2\lambda^2} = \frac{\lambda^2}{L_2\lambda\left(\lambda^2 + \omega_2^2\right)} = \frac{1}{L_2\lambda}\,\frac{\varphi^2\left(1 - \varphi_2^2\right)}{\varphi^2 - \varphi_2^2} \tag{8}$$

and clearly

$$\frac{\lambda}{Z_2} = \frac{1}{\left(\varphi^2 - \varphi_2^2\right)w_2(\varphi)}\Bigg|_{\varphi\to\varphi_2} \longrightarrow \frac{1}{L_2}\,\frac{\varphi^2\left(1 - \varphi_2^2\right)}{\varphi^2 - \varphi_2^2}\Bigg|_{\varphi\to\varphi_2}$$

which leads to the result:

$$L_2 = \varphi_2^2\left(1 - \varphi_2^2\right)w_2(\varphi_2)$$

and (9)

$$L_2C_2 = \frac{1 - \varphi_2^2}{\varphi_2^2}$$

Finally:

$$\frac{\lambda}{Z_{rem}} = \frac{\lambda}{Z_2} - \frac{\varphi^2\left(1 - \varphi_2^2\right)}{L_2\left(\varphi^2 - \varphi_2^2\right)} = \frac{1}{\varphi^2 - \varphi_2^2}\left[\frac{1}{w_2(\varphi)} - \frac{\varphi^2}{\varphi_2^2 w_2(\varphi_2)}\right] = \frac{1}{w_3(\varphi)} \tag{10}$$

In this last expression the factor $\left(\varphi^2 - \varphi_2^2\right)$ can clearly be cancelled, and the degree of the remainder is hence two less than that of $w_1(\varphi)$.

The process is repeated until the complete network is developed. The synthesis is then repeated from the output side in order to be able to check on the accuracy of the computations. In the final form of the program, the synthesis will be executed for all four impedances (input- and output-, open- and short-circuited impedances). This is done because we have found that contrary to popular beliefs, the lower degree impedances might sometimes lead to considerably better accuracy, even though they do not furnish the last branch in the ladder.

The computer finally prints the network configuration, its dual, the normalized element values and the denormalized ones for both networks. This it does four times for the four syntheses,* clearly labelling each, and always printing the first element first.

The final option available is homogeneous predistortion, with semihomogeneous predistortion ready to be incorporated but not available at this time.

One final but important point is the order in which the attenuation peaks are removed in the zero shifting process. Here three different options are available. One is the specification of this order by the designer. The other two are machine-selected specifications, the first of them automatically performed if the designer has not specified the configuration. The selection process in this option is illustrated in the following example. Let us assume that the network to be obtained is of the form of Fig. 2.

---

*In case of extreme termination ratio, there is only one synthesis.

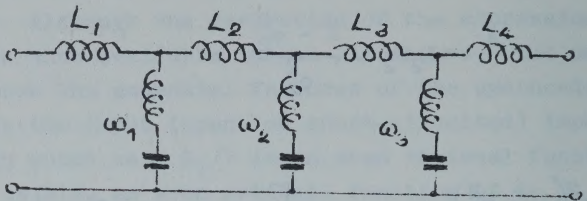


Fig. 2

The sum of the series inductances

$$L_{tot} = L_1 + L_2 + L_3 + L_4 \tag{11}$$

can be evaluated before the actual start of the synthesis. We define the average series inductance as

$$L_{ave} = \frac{1}{4} L_{tot} \tag{12}$$

The computer next evaluates $L_1^{(i)}$ for $\omega_1 = \omega^{(i)}$ where $\omega^{(i)}$ are the available attenuation peak frequencies and then selects ω_1 to be that for which L_1 is nearest to L_{ave}. The procedure is repeated at every peak removal until the whole network is developed. This process leads to a network with about the smallest spread in the values of its elements. Naturally this selection process is done only for the first synthesis, the order of the peaks is recorded and used for the other three impedance developments.

The third option is used automatically if and only if in either one of the two previous cases the computer encounters a negative element value. In such a case all previous results are dropped and the synthesis starts all over again. The selection criterion here is to pick a peak frequency which leads to the smallest positive series inductance at each step. This procedure leads to a most probably realizable network. In parenthesis we have to note, that such an arrangement of options is based upon our assumption that no apparent negative element value will be generated by loss of accuracy.

3. BAND-PASS FILTERS

The band-pass filter program is considerably more interesting and more challenging. The conventional insertion loss theory is well known. However, it is also well known that there are configurations, the general reference filter method with equal ripple type pass band, was not able to generate. The reason for this state of affairs is that these networks are not optimum in the sense that not all of their reflection coefficient zeros (zero loss points) can occur at real frequencies.[7,8] Quite recently, however, this problem was solved and now we are able to provide these configurations as well.[9,10] The favorite template method[11] for the stop-band design can easily be extended to cover these cases although the presently existing computer programs[4] do not handle these cases yet.

Our program is able to handle both the conventional as well as these new configurations, examples of which are shown in Figs. 3 and 4.

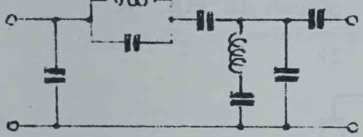


Fig. 3

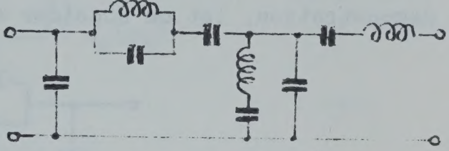


Fig. 4

The theory used for these networks is based upon Theorem 4 and a modified version of Theorem 5 of Watanabe's paper.[9] Again, both equal-ripple and maximally-flat type pass-band behaviors are available.

Turning now to the expressions, the characteristic function can be set up without difficulty.* The even polynomial in φ containing the natural modes and their negatives, is then calculated as usual and factored into factors of the form (2) just as in the low-pass case. However, instead of (3), we will associate an expression of the form

$$\left(A_i + B_i\varphi^2 + C_i\sqrt{(\beta^2-\varphi^2)(\beta^2\varphi^2-1)}\right) \qquad (13)$$

which each factor, where the coefficients are again related to the coefficients a_i and b_i through straightforward, though somewhat cumbersome expressions. We will refrain hence from listing these relationships.

---

*With the exception of those cases where Watanabe's (original or modified) Theorem 5 is used. Here, after calculating a preliminary function, the final function is obtained after an approximate cancellation of factors in numerator and denominator.

Multiplying these factors (13) together we are led to a function of the form:

$$E(\varphi) = \sqrt{(\beta^2-\varphi^2)(\beta^2\varphi^2-1)}\; E_1(\varphi) + E_2(\varphi) \qquad (14)$$

where $E_1(\varphi)$ and $E_2(\varphi)$ are both even polynomials, corresponding nevertheless to the even and odd parts of that polynomial in λ which contains the natural modes as zeros. These polynomials and the characteristic function obtained previously, again suffice for the calculation of the impedance functions, all four of them needed again for the synthesis.* The impedance expressions are again nearly identical to those in terms of λ, while the "zero shifting" is very different in form, but not in concept. We have programmed the first eight cases of Fig. 3 of Saal and Ulbrich, containing a tabulation of the possible zero shifting steps. For demonstration, let us consider case 8, shown in Fig. 5.

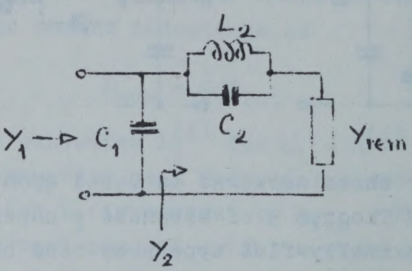


Fig. 5

Again Y_1/λ is an even rational function of λ, hence it is also an even rational function of φ:

$$Y_1/\lambda = w_1(\varphi) \qquad (15)$$

At the resonance of L_2 and C_2:

$$\lambda^2 = -\,\omega_2^2 \qquad \text{or} \qquad \varphi^2 = \frac{\beta^2 - \omega_2^2}{1 - \beta^2\omega_2^2} = \varphi_2^2 \qquad (16)$$

and here Y_1/λ is equal to C_1 giving:

---

*The termination ratio is not specified at the input in the band-pass case, since the same loss function can accommodate a wide range of termination ratios through the application of Norton's transformation. The case of extreme termination must however, be indicated since it leads to different impedance expressions and only one synthesis.

$$C_1 = Y_1/\lambda \Big|_{\varphi=\varphi_2} = w_1(\varphi_2) \tag{17}$$

Consequently:

$$Y_2/\lambda = Y_1/\lambda - C_1 = w_1(\varphi) - w_1(\varphi_2) = \left(\varphi^2 - \varphi_2^2\right) w_2(\varphi) \tag{18}$$

because $w_1(\varphi)$ is an even function.

The impedance of the resonant circuit is given by:

$$Z_{res} = \frac{L_2\lambda}{1 + L_2C_2\lambda^2} = \frac{1}{C_2\lambda}\,\frac{\left(1 - \beta^2\varphi_2^2\right)(\beta^2-\varphi^2)}{(\beta^4-1)\left(\varphi^2 - \varphi_2^2\right)} \tag{19}$$

using Eq. (1b). Again as $\varphi \to \varphi_2$ we see, that $Y_2/\lambda \to 1/\lambda Z_{res}$ and this gives:

$$C_2 = \frac{\left(1 - \beta^2\varphi_2^2\right)\left(\beta^2 - \varphi_2^2\right)}{(\beta^4 - 1)}\, w_2(\varphi_2)$$

with (20)

$$L_2C_2 = \frac{1 - \beta^2\varphi_2^2}{\beta^2 - \varphi_2^2}$$

The remainder is then obtainable as:

$$\frac{\lambda}{Y_{rem}} = \frac{\lambda}{Y_2} - \lambda Z_{res} = \frac{1}{\varphi^2 - \varphi_2^2}\left[\frac{1}{w_2(\varphi)} - \frac{\beta^2 - \varphi^2}{\beta^2 - \varphi_2^2}\;\frac{1}{w_2(\varphi_2)}\right] = \frac{1}{w_3(\varphi)} \tag{21}$$

where the factor $\left(\varphi^2 - \varphi_2^2\right)$ is again cancelled.

All other types of sections can be similarly dealt with.

The final problem of the synthesis is again the configuration, which has considerably more variations than in the low-pass case. To render the problem more tractable we separated the problem of the configuration proper from that of the exact order of loss peaks. (We might recall, that in the low-pass case the configuration meant only the exact order of the peaks, apart from the dual network.) Clearly the same configuration can realize the same set of loss peaks in several different orders, while the same order of peaks can be realized in several different configurations. Therefore, the exact specification will require us to prescribe both. The process is further complicated by the fact that the configuration specification in the reverse direction is not simply the inverse of the forward specification and the former might not exist even though the latter does. (See below.)

The available options are as follows:

1. Specification of both configuration and order of loss peaks by designer. This is achieved by specifying what operations, listed in Fig. 3 of Saal and Ulbrich's paper, are to be performed sequentially, and at each step which one of the loss peaks, if any, is to be removed.

2. Specification of the configuration only, by the designer. Here the sequence of operations is listed as before, but the actual loss peak frequency is selected from the list of available values by the computer. The aim here is to obtain a network which is most probably realizable and the procedure to be used is analogous to that described in the low-pass case. The method is simply to select the peak at each step, which requires the minimum immittance in the zero shifting.

3. Specification of both by the computer. This is done in two steps; first a subroutine determines a configuration, then the order of the loss peaks is determined as in case 2 above. Due to the large variety of configurations possible, it was necessary to restrict its machine specification to a specific type. For obvious reasons the minimum inductance version was selected, in such a form that the same network could be synthesized from both ends. This restriction enabled us to give a useful and unique prescription for the configuration specification. The sole exception is the case where no minimum inductance structure exists, in which case the subroutine will state this fact and terminate the synthesis.

The structure determined by the computer starts, after a possible preliminary step, with an alternating series of finite attenuation peaks from the upper and lower stop bands, until at least one group is exhausted. It then removes all the remaining finite peaks and ends by removing all the remaining extreme attenuation peaks. Finally the subroutine properly inverts the sequence of orders to describe the same network for the reverse synthesis.

As an example, assume that we have a network with one finite peak below and 3 finite peaks above the pass band, a single peak at zero frequency and a triple peak at infinity. The network configuration selected by the computer will be that of Fig. 6.

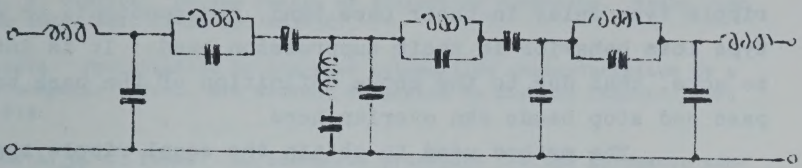


Fig. 6

The network is, in this case, completely unique, apart from the exact order of the peaks above the band; and can be synthesized from both directions.

That this last feature is not obvious, is illustrated by the equivalent configuration below, which cannot be synthesized in that form from both directions:

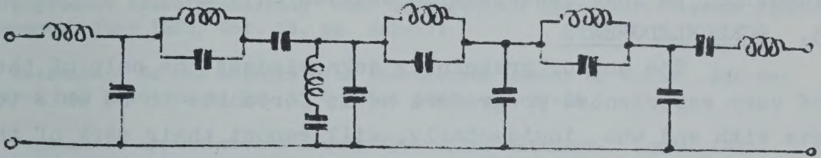


Fig. 7

While it is clear that Fig. 7 can be converted into the form of Fig. 6 by repeated application of Norton's transformation, it is preferable to be able to avoid such transformations at this stage, even though they will almost certainly become necessary later to accommodate the required termination ratio. The less one has to manipulate the filter the less chances there are of introducing errors, especially since these transformations will still have to be done manually.

Finally we might mention that contrary to the low-pass case, the dual configuration is not printed in the output, since in most cases the minimum inductance structure is preferred, the dual of which would be the uninteresting minimum capacitance structure.

Credit must be given here to a group of Japanese specialists, who have developed a band-pass synthesis method very similar to ours.[12] At present, sufficient information about their method is not available to enable us to make a meaningful comparison.

4. CONSTANT-DELAY NETWORKS

The last program in the group presently under development is a constant-delay low-pass filter synthesis program. This provides us with low-pass filters having maximally flat or equal ripple type delay in their pass band, and monotonic or equal minima type loss behavior in their suppression band. It is interesting to note, that due to the above definition of the pass band, the pass and stop bands can overlap here.

The method used to obtain the equal ripple type delay behavior is that described by Ulbrich and Piloty,[13] with the properly scaled maximally flat delay function as starting point. We might point out the fact that the transformed variable used by Ulbrich and Piloty is essentially the same as ours. The method to obtain the equal minima type stop band behavior is described by the author elsewhere.[14]

After all functions are computed, the actual synthesis is identical to that described in Section 2.

5. ACKNOWLEDGMENTS

The author gratefully acknowledges the help of the group of very experienced programers he is fortunate to be able to cooperate with and who, incidentally, will report their part of the story at a later date. The close cooperation necessary in an undertaking of this size, influenced the author's ideas considerably and beneficially.

6. References

1. S. Darlington: Synthesis of reactance fourpoles which produce prescribed insertion loss characteristics. Journ. of Mathematics and Physics. September 1939 vol. 18, pp. 257-353.

2. R. Saal, E. Ulbrich: On the design of filters by synthesis. Trans. IRE December 1958, Vol. CT-5, pp. 287-327.

3. T. Fujisawa: Optimization of low-pass attenuation characteristics by a digital computer. Proc. 6th Midwest Symposium on Circuit Theory. 1963 pp. P1-P13.

4. B. R. Smith, G. C. Temes: An iterative procedure for the automatic synthesis of general parameter low-pass and band-pass filters. Unpublished memorandum. July 1963.

5. F. S. Farkas, F. J. Hallenbeck, F. E. Stehlik: Band-pass filter, band elimination filter and phase simulating network for carrier program systems. Bell System Tech. Journ., April 1949, Vol. 28, pp. 196-220.

6. J. E. Colin: Apercus sur une theorie des quadripoles en echelle passe-bas et passe-haut. Cables et Transmission. No. 2, 1959, Vol. 13, pp. 75-93.

7. G. Gotz: Die Struktur der Kettenmatrix von Reaktanzvierpolen bei vorgegebener Abzweigschaltung oder vorgegebener charakteristischer Funktion. Frequenz. June 1959, Vol. 13, pp. 166-171.

8. H. Watanabe: On the network with the minimum number of coils. NEC Research and Development No. 1, October 1960, pp. 26-35.

9. H. Watanabe: Approximation theory for filter networks. Trans. IRE September 1961, Vol. CT-9, pp. 341-356.

10. J. E. Colin: Un nouveau filtre symetrique de synthese en echelle: le filtre passe-bande parametrique. Cables et Transmission. No. 1, January 1962, Vol. 16, pp. 31-40.

11. E. Rumpelt: Schablonenverfahren fur den Entwurf elektrischer Wellenfilter auf der Grundlage der Wellenparameter. Telegraphen-, Fernsprech-, Funk- und Fernseh-Technik. August 1942. Vol. 31, pp. 203-210.

12. T. Iedokoro, T. Tsuchiya, H. Watanabe: A new calculation method for the design of filters by digital computer with the special consideration of the accuracy problem. IEEE Internat. Conv. Rec. 1963 Vol. 11, Part 2, pp. 100-112.

13. E. Ulbrich, H. Piloty: Uber den Entwurf von Allpassen, Tiefpassen und Bandpassen mit einer im Tschebyscheffschen Sinne approximierten konstanten Gruppenlaufzeit. Archiv der elektrischen Ubertragung. October 1960, Vol. 14, pp. 451-467.

14. G. Szentirmai: The design of arithmetically symmetrical band-pass filters. Ph.D. Thesis, Polytechnic Institute of Brooklyn, Brooklyn 1, New York, June 1963.

A NOTE ON ZERO-STATE STABILITY OF LINEAR SYSTEMS*

C. A. DESOER and A. J. THOMASIAN
Department of Electrical Engineering and
Electronic Research Laboratory
University of California
Berkeley, California

We consider a linear, time-varying, possibly anticipative system whose zero-state response y to an input x is given by

$$y(t) = \int_{-\infty}^{\infty} w(t, \tau)\, x(\tau)\, d\tau \qquad -\infty < t < \infty \qquad (1)$$

For simplicity we assume that all functions are real-valued. We will consider only bounded inputs and we require that the response always be defined and finite for any such input. Since we take the response to be defined by (1), we consider exclusively the zero-state response of the system.† Now, for any fixed t_o, we can define a bounded input by

$$x_{t_o}(\tau) = \operatorname{sgn} w(t_o, \tau) \qquad -\infty < \tau < \infty \qquad (2)$$

and for this input the response at t_o is

$$\int_{-\infty}^{\infty} \left| w(t_o, \tau) \right| d\tau \qquad (3)$$

Therefore our requirement implies that the integral (3) is finite for each t_o. Thus we are led to our basic assumption concerning w, which characterizes the system:

Assumption A: The linear time-varying system defined by (1) satisfies the condition that for every fixed t_o, $-\infty < t_o < \infty$, $w(t_o, \tau)$ is a measurable function of τ and the integral (3) is finite.

Now let $\mathcal{Y}$ be the set of all functions on $(-\infty, \infty)$ and if a function y happens to be bounded, define its norm by

$$\|y\| = \sup |y(t)| \qquad -\infty < t < \infty \qquad (4)$$

Let $\mathcal{X}$ be the set of bounded measurable functions on $(-\infty, \infty)$. It is well known that the space $\mathcal{X}$ with the norm (4) is complete, therefore is a Banach space. All of our inputs x will be long to $\mathcal{X}$, and so have finite norm ($\|x\| < \infty$)

---

* This research is supported in part by the National Science Foundation, under Grant G-15965.

† The zero-state response of a linear system is the response of the system when it starts from the zero state. For further information on this subject, consult Zadeh, L. A. and C. A. Desoer: Theory of Linear Systems, The state-space approach, McGraw-Hill Book Co; New York, 1963.

while, because of Assumption A, all of our responses y are well defined (the integral in (1) is taken as a Lebesque integral), finite (y(t) is finite for each t, $-\infty < t < \infty$), and belong to $\mathcal{Y}$. Thus under Assumption A, formula (1) defines a linear operator T from $\mathcal{X}$ to $\mathcal{Y}$, so that $y = Tx$.

Now probably the most natural definition of "zero-state stability" in the sense of, "any bounded input produces a bounded output," is given by Property 1 below, while Property 2 is an apparently stronger requirement. Property 3 is a common hypothesis for a linear system. We list now these three properties of interest to us, which a linear system satisfying Assumption A might have:

Property 1. For all x belonging to $\mathcal{X}$, the output $y = Tx$ is bounded, i.e. x belongs to $\mathcal{X}$ (so $\|x\| < \infty$) implies $\|Tx\| < \infty$.

Property 2. Property 1 holds and furthermore there is a finite constant C such that $\|Tx\| \leq C\|x\|$ for all x belonging to $\mathcal{X}$.

Property 3. There is a finite constant M such that

$$\int_{-\infty}^{\infty} |w(t, \tau)|\, d\tau < M \text{ for all } t, \ -\infty < t < \infty.$$

Thus Property 1 requires that $\|Tx\| / \|x\|$ be finite for every $x \neq 0$ belonging to $\mathcal{X}$, while Property 2 requires that all of these ratios be bounded by a constant C which is independent of x. The interesting point is that all these properties are equivalent.

Theorem. A linear system satisfying Assumption A either has all three of the above properties or none of them.*

Proof. We will show $3 \Longrightarrow 2 \Longrightarrow 1 \Longrightarrow 3$. $2 \Longrightarrow 1$ is immediate.

Proof that $3 \Longrightarrow 2$. For any $x \in \mathcal{X}$ we have†

$$\|Tx\| = \sup_t \left| \int (w(t, \tau)\, x(\tau)\, d\tau \right| \leq \sup_t \int |w(t, \tau)|\, |x(\tau)|\, d\tau$$

$$\leq \sup_t \left(\|x\| \int |(w(t, \tau)|\, d\tau \right) \leq M \|x\|$$

Proof that $1 \Longrightarrow 3$. The key to the proof is the "Principle of Uniform Boundedness,"** which asserts: if for each t in some index set R, T_t is a bounded linear operator mapping a Banach space $\mathcal{X}$ into a Banach space $\mathcal{Y}$, and if, for every $x \in \mathcal{X}$, the set of real numbers $\{\|T_t(x)\| : t \in R\}$ is bounded, then the set of numbers $\{\|T_t\| : t \in R\}$ is bounded.

---

* A constructive proof of this statement has been given by D. C. Youla, "On the Stability of Linear Systems", IEEE Transactions on Circuit Theory, vol. CT-10, no. 2, pp. 276-279, June 1963.

† All suprema and all integrals are taken over $(-\infty, \infty)$.

**See for example, p. 168 of C. Goffman's article in Buck, R.C., "Studies in Modern Analysis," published by the Mathematical Association of America, Distributed by Prentice Hall, 1962. See also Dunford, N. and Schwartz, J. T. Linear Operators, Part I. Interscience Publishers, 1958, Theorem II, p. 52.

Now, for each $t \in R = (-\infty, \infty)$, (1) defines a linear operator from the Banach space $\mathcal{X}$ into the Banach space $\mathcal{Z} = (-\infty, \infty)$ (with absolute value norm) by $T_t(x) = y(t)$. Clearly $\|T_t\| = \sup\limits_{\|x\|=1} \left\{ \|T_t(x)\| \le \int |w(t, \tau)| d\tau \right\}$ while for the particular x given by (2) we have $\|T_{t_o}\| \ge \int |w(t_o, \tau)| d\tau$ so $\|T_t\| = \int |w(t, \tau)| d\tau < \infty$ for all t. Property 1 states that for all $x \in \mathcal{X}$ $\sup\limits_{t \in R} \|T_t(x)\| = \sup\limits_{t \in R} \|y(t)\| = \|T(x)\| < \infty$. Hence by the conclusion of the Principle of Uniform Boundedness, Property 3 follows.

RESPONSE FUNCTIONS OF LINEAR TIME VARYING SYSTEMS

J. C. LINDENLAUB
Lincoln Laboratory*
Massachusetts Institute of Technology
Cambridge, Massachusetts
(Assistant Professor, School of Electrical Engineering
Purdue University, Lafayette, Indiana)

ABSTRACT

Fourier transforms of the impulse response of time varying systems have been treated by several workers.** This paper places an emphasis on the physical interpretation that can be given to these transforms. A particular set of response functions which use h(t, τ), the response at time t due to an impulse applied at time τ, as a starting point are shown to have the physical interpretation of impulse response, since wave response, impulse spectrum, and sine wave spectrum. The mathematical assumptions necessary to insure the existence of these two dimensional Fourier transforms are carefully stated. By introducing a general linear transformation on the t and τ variables a new set of response functions can be generated. It is shown however that in only very special cases can simple physical interpretations be given to these new response functions.

Introduction

Much of the earlier work on the subject of varying parameter systems has dealt with the solution of differential equations with time dependent coefficients and the relationship of the differential equation to the impulse response. These topics, as well as other contributions, have been reviewed by Zadeh.[1] Zadeh's paper also contains an extensive bibliography on the subject of time varying linear systems.

This paper treats linear time varying networks from the systems viewpoint where the emphasis is primarily on the relationships between the input and output signals. This is the viewpoint taken by the radar or communications engineer who treats the radar target configuration or communication channel as a black box which performs a linear time dependent transformation on the input signal to yield the output signal.

A Set of Time-Varying Response Functions

A set of functions is defined in this section that may be used to relate the time or frequency domain response of linear time varying systems to the time or frequency representation of the excitation. The notation used is summarized in Fig. 1. Throughout this report ω will be used to designate the radian frequency of the input signal and ν will be used to denote the frequency of the output signal.

The response at time t to an impulse applied at time τ will be called the impulse response, h(t, τ), of the time varying system. The variable t is the instant of observation while τ is the instant at which the impulse is applied. Only physically realizable systems will be considered so that $h(t, \tau) = 0$ for $t < \tau$ and, since $h(t, \tau)$ is the response of a physical system to the real excitation $\delta(t - \tau)$, $h(t, \tau)$ is a real function of t and τ. In addition, it is assumed that $h(t, \tau)$ is square summable, i.e.,*

---

* Operated with support from the U.S. Advanced Research Projects Agency.

**See for instance the bibliography in the paper by Zadeh referenced in the body of this paper.

$$h(t, \tau)^2 \, dt d\tau < \infty.$$

Note that this property is preserved under a linear transformation of the variables t and τ. This assumption is stronger than need be for some of the results that follow, but it is mathematically convenient and not totally unrealistic from a physical standpoint. Essentially, it requires the impulse response to have finite energy (the t integration), and, in addition, the response must approach zero sufficiently fast as the application time approaches $-\infty$, or $+\infty$ (the τ integration). A time invariant filter, where $h(t, \tau)$ is a function only of $t - \tau$, does not satisfy this condition. Formally, however, it may be treated as a singular case. Alternatively, one might argue that the system or experiment was turned on or started sometime in the finite past and that the system or experiment will be turned off or terminated sometime in the finite future.

The square summable assumption allows the definition of a set of Fourier transforms of $h(t, \tau)$ to be made; each transform is also square summable.[2] The particular set of transforms used here is summarized in Fig. 2. Note that transforming from the frequency variable to the time variable always involves division by 2π, but that the signs on the exponential term in the $\tau - \omega$ transformation is opposite to that used in the $t - \nu$ transformation. This seemingly inconsistent notation is used because it leads to simple physical interpretations for h^T, H, and H^T.

From the fact that $h(t, \tau)$ is a real function of t and τ and the definitions of h^T, H, and H^T, it follows that

$$\begin{aligned} H^*(t, \omega) &= H(t, -\omega) \\ h^{T*}(\nu, \tau) &= h^T(-\nu, \tau) \qquad (1\text{-}1) \\ H^{T*}(\nu, \omega) &= H^T(-\nu, -\omega). \end{aligned}$$

To obtain a physical interpretation of $H(t, \omega)$, the superposition property of linear systems is used to state that the time response to an arbitrary input $x(t)$ is

$$y(t) = \quad x(\tau)h(t, \tau)d\tau. \qquad (1\text{-}2)$$

If, formally, the special case of $x(t) = e^{i\omega t}$ is now considered and substituted in Eq. (1-2), one obtains

---

* Integrals without specified limits will be taken to mean integration over the range $-\infty$, ∞.

$$y(t)_{x(t) = e^{i\omega t}} = H(t, \omega) = \text{the time response at time t to the excitation } e^{i\omega t} \qquad (1\text{-}3)$$

H(t, ω) will be termed the sinewave response of the time-varying system. It is often convenient to represent x(t) in terms of its Fourier transform

$$x(t) = \frac{1}{2\pi} \cdot \quad X(\omega)\, e^{i\omega t}\, d\omega \qquad (1\text{-}4)$$

where

$$X(\omega) = \quad x(t)\, e^{-i\omega t}\, dt. \qquad (1\text{-}5)$$

With Eq. (1-4) substituted in Eq. (1-2), there results

$$y(t) = \frac{1}{2\pi} \qquad X(\omega) h(t, \tau)\, e^{i\omega t}\, d\tau d\omega. \qquad (1\text{-}6)$$

The τ integration in Eq. (1-6) may be carried out to yield

$$y(t) = \frac{1}{2\pi} \quad X(\omega) H(t, \omega) d\omega. \qquad (1\text{-}7)$$

Equation (1-4) may be interpreted as representing x(t) as a superposition of $e^{i\omega t}$ functions, so that Eq. (1-7) may be viewed as expressing y(t), as a superposition of sinewave responses.

The frequency response to an excitation expressed in the time domain as x(t) will now be considered. Starting with Eq. (1-2) and transforming the t variable, there results

$$Y(\nu) = \quad x(\tau) h(t, \tau)\, e^{-i\nu t}\, d\tau dt$$

$$Y(\nu) = \quad x(\tau) h^T(\nu, \tau) d\tau. \qquad (1\text{-}8)$$

A physical interpretation of $h^T(\nu, \tau)$ may be obtained by letting x(t) be an impulse; thus $h^T(\nu, \tau)$ is the frequency response to an impulse applied at time τ. For later reference $h^T(\nu, \tau)$ will be termed the impulse spectrum.

Now consider the frequency response to an excitation expressed in the frequency domain as X(ω). From Eq. (1-7)

$$Y(\nu) = \frac{1}{2\pi} \quad X(\omega) H(t, \omega)\, e^{-i\nu t}\, d\omega dt. \qquad (1\text{-}9)$$

Upon carrying out the t integration

$$Y(\nu) = \frac{1}{2\pi} \quad X(\omega) H^T(\nu, \omega) d\omega. \qquad (1\text{-}10)$$

The interpretation of $H^T(\nu, \omega)$ becomes apparent if the spectrum for $e^{i\omega t}$ is substituted in Eq. (1-10). Thus, $H^T(\nu, \omega)$ is physically the frequency response to an excitation $e^{i\omega t}$. Hence, $H^T(\nu, \omega)$ shall be called the sinewave spectrum.

In summary, the basic relationships between the input and output of a time-variant system are

$$y(t) = \quad x(\tau)h(t, \tau)d\tau \tag{1-11}$$

$$y(t) = \frac{1}{2\pi} \quad X(\omega)H(t, \omega)d\omega \tag{1-12}$$

$$Y(\nu) = \quad x(\tau)h^{T}(\nu, \tau)d\tau \tag{1-13}$$

$$Y(\nu) = \frac{1}{2\pi} \quad X(\omega)H^{T}(\nu, \omega)d\omega. \tag{1-14}$$

Some properties of h^T, H, and H^T which result from the physical realizability and square summable assumptions on $h(t, \tau)$ will now be discussed. First, it is useful to state that

$$h(t, \tau)\ ^{2}\ dt < \infty \text{ for almost all } \tau \tag{1-15}$$

$$h(t, \tau)\ ^{2}\ dt < \infty \text{ for almost all } t. \tag{1-16}$$

These conditions plus the fact that physical realizability requires $h(t, \tau)$ to vanish on a half-line lead to the facts that $H(t, \omega)$ is analytic in the lower half ω-plane for almost all t and $h^T(\nu, \tau)$ is analytic in the lower half ν-plane for almost all τ.

The analytic behavior of $H(\nu, \omega)$ is most easily handled by introducing the change variable

$$H^{T}(\nu, \omega) = H'^{T}\left(\frac{\omega + \nu}{2}, \omega - \nu\right) = H'^{T}(\mu, \eta). \tag{1-17}$$

The corresponding change of variable in the time domain is

$$h(t, \tau) = h'\left(t - \tau, \frac{t + \tau}{2}\right) = h'(r, s). \tag{1-18}$$

The invariance of the square summable property under a linear transformation of variables allows one to state

$$H'(r, \eta)\ ^{2}\ dr < \infty \text{ for almost all } \eta \tag{1-19}$$

where

$$H'(r, \eta) = \quad h'(r, s)\ e^{i\eta s}\ ds = \frac{1}{2\pi} \quad H'^{T}(\mu, \eta)\ e^{i\mu r}\ d\mu. \tag{1-20}$$

From this fact and the realizability condition $h'(r, s) = 0$ for $r < 0$, it can be shown that

$$H'^{T}(\mu, \eta) = H'^{T}\left(\frac{\omega + \nu}{2}, \omega - \nu\right) \tag{1-21}$$

is analytic in the lower half μ-plane for almost all η.

Linear Transformations of the Response Functions

The function $h(t, \tau)$ as defined above is a particularly convenient manner to specify the impulse response because of the physical significance that can be attached to the various Fourier transforms of $h(t, \tau)$ In formulating circuit models of time varying systems and in deriving the response functions of particular channels from the physics involved, it is often convenient to express the impulse response function in terms of the variables r and s which are derived from t and τ by the linear transformation

$$\begin{array}{c} r \\ \\ s \end{array} = \begin{array}{cc} a & b \\ & \\ c & d \end{array} \cdot \begin{array}{c} t \\ \\ \tau \end{array} \tag{2-1}$$

and the relation

$$h'(r, s) = h(t, \tau). \tag{2-2}$$

Consideration of transforms of this type also makes it possible to relate the work presented here to that presented by earlier workers in the field.

It was pointed out earlier that the square summable property of $h(t, \tau)$ is invariant under this type of coordinate transformation, thus it is possible to define a set of Fourier transformations

$$h'^{T}(\mu, s) = h'(r, s)\, e^{-i\mu r}\, dr \tag{2-3}$$

$$H'(r, \eta) = h'(r, s)\, e^{i\eta s}\, ds \tag{2-4}$$

$$H'^{T}(\mu, \eta) = h'(r, s)\, e^{i\eta s - i\mu r}\, drds. \tag{2-5}$$

These transforms and the inverse relationships are summarized in Fig. 3. It will become evident from the discussion below that the input-output expressions Eq (1-11) and Eq. (1-14) may be expressed as

$$y(t) = h'(at + b\tau, ct + d\tau)\, x(\tau)\, d\tau \tag{2-6}$$

and

$$Y(\nu) = \frac{\Delta}{2\pi} H'^{T}\left(\frac{d\nu + c\omega}{\Delta}, \frac{b\nu + a\omega}{\Delta}\right) X(\omega) d\omega \tag{2-7}$$

where Δ is the determinant of the transformation matrix in Eq. (2-1). It will also become clear that Eq. (1-12) and Eq. (1-13) are not always directly expressible in terms of h' and its transforms.

It is natural to inquire about the relationships which exist between the h circle of transforms and the h' circle of transforms. The linear transformation Eq. (2-1) and the identity Eq. (2-2) can always be combined to relate h and h' as

$$h(t, \tau) = h'(at + b\tau, ct + d\tau) \tag{2-8}$$

or

$$h'(r, s) = h\left(\frac{dr - bs}{\Delta}, \frac{-cr + as}{\Delta}\right). \tag{2-9}$$

From this second relationship the physical significance of h'(r, s) can be stated in the following manner. h'(r, s) is the response at time $\frac{dr - bs}{\Delta}$ to an impulse applied at time $\frac{as - cr}{\Delta}$. This interpretation becomes more meaningful when specific examples are considered. For instance, when $r = t - \tau$ and $s = t(\Delta = 1)$ h'(r, s) takes on the meaning of the response observed at time s due to an impulse applied at time s-r. The variable r is readily identified as the "age" of the input. s is the observation time, but the s dependence of h'(r, s) in this case may also be interpreted as describing the time varying nature of the system in terms of the observation time.

The relationships between the sinewave spectrum and H'^{T} may be derived in the following manner

$$H^{T}(\nu, \omega) = h(t,\tau)\, e^{i\omega\tau - i\nu t}\, dt d\tau. \tag{2-10}$$

Expressing t and τ in terms of r and s yields

$$H^{T}(\nu, \omega) = \Delta \quad h'(r, s)\, e^{i\omega\left(\frac{as - cr}{\Delta}\right)}\, e^{-i\nu\left(\frac{dr - bs}{\Delta}\right)}\, drds$$

$$= \Delta \; H'^{T}\left(\frac{d\nu + c\omega}{\Delta}, \frac{b\nu + a\omega}{\Delta}\right) \tag{2-11}$$

or

$$H^{T}(\nu, \omega) = \Delta \; H'^{T}(\mu, \eta) \tag{2-12}$$

where μ and η are related to ν and ω by the transformation

$$\begin{matrix} \mu \\ \\ \eta \end{matrix} = \frac{1}{\Delta} \begin{matrix} d & c \\ & \\ b & a \end{matrix} \quad \begin{matrix} \nu \\ \\ \omega \end{matrix} \tag{2-13}$$

Inverting Eq. (2-11) results in

$$H'^{T}(\mu, \eta) = \frac{1}{\Delta} H^{T}(a\mu - c\eta, \; -c\mu + d\eta). \tag{2-14}$$

This last expression allows $H'^{T}(\mu, \eta)$ to be interpreted as the output spectrum, as a function of $a\mu - c\eta$, resulting from a sinewave input of frequency $d\eta - c\mu$. For the example cited above ($r = t - \tau$, $s = t$)

$$H'^{T}(\mu, \eta) = H^{T}(\mu - \eta, \mu) \tag{2-15}$$

so that the frequency μ can be identified with the input frequency and η is the difference between the input frequency and the output frequency. Thus η is a measure of the frequency translation or band spreading caused by the time varying nature of the system.

Only when one of the elements a, b, c, or d of the linear transformation is equal to zero can h^{T} (or H but not both) be expressed directly in terms of h'^{T} or H'.* A zero in the transformation matrix results in one of the t or τ time domain variables being proportional to one of the r or s time domain variables (and one of the ν or ω frequency domain variables is proportional to one of the μ or η frequency domain variables). This proportionality relationship allows one to transform h' with respect to one of the t or τ (but not both) variables.

Considering the $r = t - \tau$, $s = t$ example once again, $H(t, \omega)$ can be expressed as

$$H(t, \omega) = \quad h'(t - \tau, t)\, e^{i\omega\tau}\, d\tau = h'^{T}(\omega, t)\, e^{i\omega t} . \qquad (2\text{-}16)$$

Equation (2-16) allows one to relate $h'^{T}(\omega, t)$ to the sinewave response. Formally, $h'^{T}(\omega, t)$ is simply the response to $e^{i\omega t}$ divided by $e^{i\omega t}$. It is worthwhile pointing out the difficulty encountered when one attempts to express $h^{T}(\nu, \tau)$ in terms of the primed functions.

$$h^{T}(\nu, \tau) = \quad h'(t - \tau, t)\, e^{-i\nu t}\, dt. \qquad (2\text{-}17)$$

The appearance of t in both arguments of h' prevents one from expressing the integral in terms of the Fourier transforms of h'. For this same reason there does not appear to be a simply physical interpretation for $h'(r, s)$ in this case.

The relationships between the primed and unprimed functions are pictured in Fig. 4. The solid lines indicate that Fourier transform relationships exist between the connected functions while dashed lines indicate the existence of linear transformation relationships. While in general there is no direct relationship between say h'^{T} and H one can always get from h'^{T} to H by the series of transforms $h'^{T} \quad H'^{T}$, $H'^{T} \quad H^{T}$, and $H^{T} \quad H$.

Linear transformations can be used to relate the time varying response functions introduced in this work to those presented in earlier literature. Zadeh[1] has defined a time varying system function as

$$h(t, \tau)\, e^{-i\omega(t - \tau)}\, d\tau.$$

Writing

$$h(t, \tau) = h'(r, s) = h'(t - \tau, t), \qquad (2\text{-}18)$$

---

* Linear transformations with a = d = 0 or b = c = 0 allow all of the unprimed (h. etc.) funtions to be expressed in terms of the primed (h', etc.) functions but these "more trivial cases" will not be pursued here.

it becomes evident that Zadeh's system functions is equal to $h'^{T}(\omega, t)$. The relationship of this function to $H(t, \omega)$ was discussed after Eq. (2-16). Zadeh's bifrequency function is identical to $H^{T}(\nu, \omega)$.

Several impulse response functions have been introduced by Kailath.[3] The relationships between Kailath's notation and the $h(t, \tau)$ used here are given by

$$h_1(t, \tau) = h(t, \tau) \tag{2-19}$$

$$h_2(z, \tau) = h(t - \tau, \tau) \tag{2-20}$$

$$h_3(y, t) = h(t - \tau, t). \tag{2-21}$$

Hence Hailath's $h_1(t, \tau)$ is identical to $h(t, \tau)$, $h_3(y, t)$ is, by Eq. (2-9), the response at time t to an impulse applied y seconds ago and $h_2(z, \tau)$ is the response at time $z + \tau$ to an impulse applied at time τ.

Concluding Remarks

This paper has shown that the impulse response function $h(t, \tau)$ as defined above leads to a circle of Fourier transforms in which each member has a satisfying physical interpretation. Linear transformation of the t and τ variables allow one to relate other definitions of the impulse response (and the corresponding circle of transforms) to $h(t, \tau)$(and its circle of transforms).

In another report[4] the author has treated a number of related topics in the context of the response functions presented here. Among them are cascading of time-varying systems, bandpass systems, separable response functions, and general models of time-varying systems.

Acknowledgments

The author wishes to express his appreciation to Dr. E. J. Kelly for the many fruitful discussions with him throughout this effort and to Dr. R. T. Prosser for his comments concerning the mathematical properties of the response functions.

References

1. Zadeh, L. A., "Time Varying Networks, I," Proc. IRE, Vol. 49 October 1961, pp. 1488-1503.
2. Titchmarsh, E. C., Theory of Fourier Integrals, Oxford Clarendon Press, 1948, p. 69.
3. Kailath, T., "Channel Characterization: Time-Variant Dispersive Channels," Chapter 6 of Lectures on Communication System Theory, E. J. Baghdady, editor McGraw-Hill, 1961.
4. Lindenlaub, J. C., "Response Functions of Linear Time Varying Systems," M.I.T. Lincoln Laboratory Group Report 41G-6, August 1963.

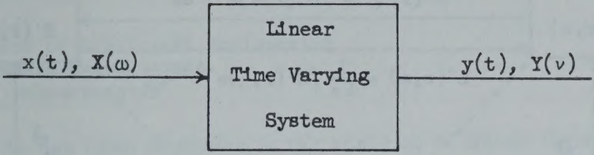

$x(t),\ X(\omega)$
Linear
Time Varying
System
$y(t),\ Y(\nu)$

Fig. 1

Input-Output Notation

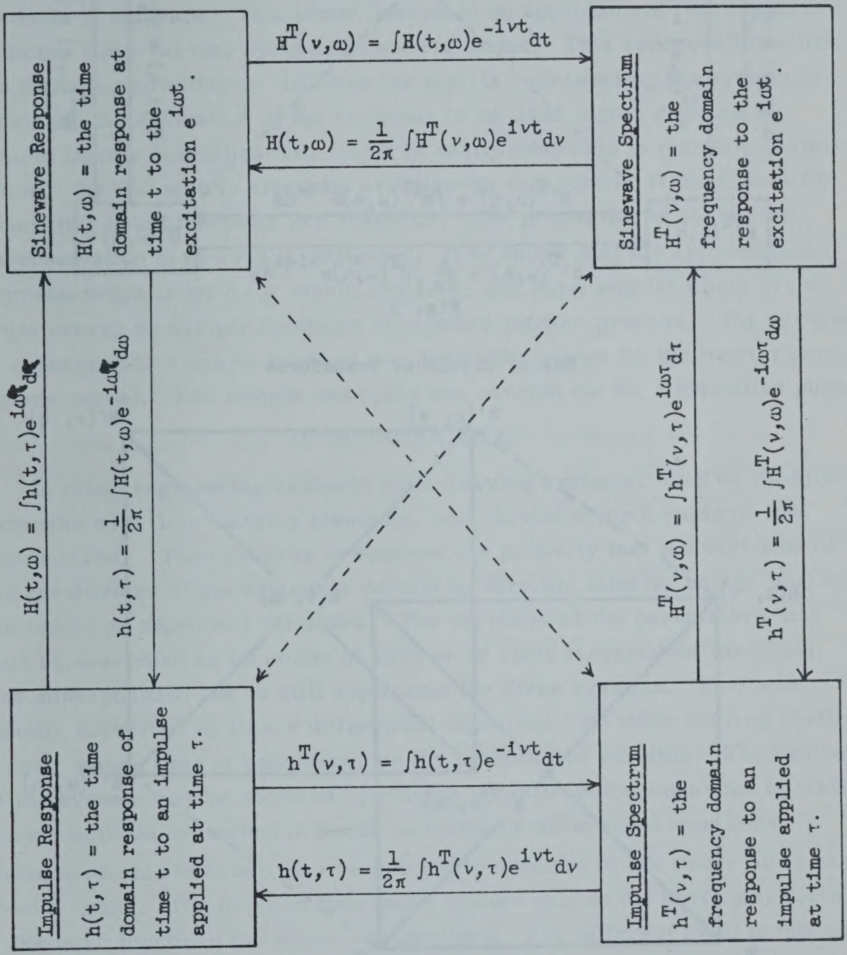

Sinewave Response
$H(t,\omega)$ = the time domain response at time t to the excitation $e^{i\omega t}$.
Sinewave Spectrum
$H^T(\nu,\omega)$ = the frequency domain response to the excitation $e^{i\omega t}$.
Impulse Response
$h(t,\tau)$ = the time domain response of time t to an impulse applied at time τ.
Impulse Spectrum
$h^T(\nu,\tau)$ = the frequency domain response to an impulse applied at time τ.
$H^T(\nu,\omega) = \int H(t,\omega)e^{-i\nu t}dt$
$H(t,\omega) = \frac{1}{2\pi}\int H^T(\nu,\omega)e^{i\nu t}d\nu$
$h^T(\nu,\tau) = \int h(t,\tau)e^{-i\nu t}dt$
$h(t,\tau) = \frac{1}{2\pi}\int h^T(\nu,\tau)e^{i\nu t}d\nu$
$H(t,\omega) = \int h(t,\tau)e^{i\omega\tau}d\tau$
$h(t,\tau) = \frac{1}{2\pi}\int H(t,\omega)e^{-i\omega\tau}d\omega$
$H^T(\nu,\omega) = \int h^T(\nu,\tau)e^{i\omega\tau}d\tau$
$h^T(\nu,\tau) = \frac{1}{2\pi}\int H^T(\nu,\omega)e^{-i\omega\tau}d\omega$

Fig. 2

Response Function Fourier Transform Relationships

$$\begin{tikzcd}
h'(r,s) \arrow[r, shift left, "{H'(r,\eta) = \int h'(r,s)e^{i\eta s}ds}"] \arrow[d, shift left, "{h'^T(\mu,s) = \int h'(r,s)e^{-i\mu r}dr}"] \arrow[dr, dash, dashed] & H'(r,\eta) \arrow[l, shift left, "{h'(r,s) = \frac{1}{2\pi}\int H'(r,\eta)e^{-i\eta s}d\eta}"] \arrow[d, shift left, "{H'^T(\mu,\eta) = \int H'(r,\eta)e^{-i\mu r}dr}"] \arrow[dl, dash, dashed] \\
h'^T(\mu,s) \arrow[u, shift left, "{h'(r,s) = \frac{1}{2\pi}\int h'^T(\mu,s)e^{i\mu r}d\mu}"] \arrow[r, shift left, "{H'^T(\mu,\eta) = \int h'^T(\mu,s)e^{i\eta s}ds}"] & H^T(\mu,\eta) \arrow[l, shift left, "{h'^T(\mu,s) = \frac{1}{2\pi}\int H^T(\mu,\eta)e^{-i\eta s}d\eta}"] \arrow[u, shift left, "{H'(r,\eta) = \frac{1}{2\pi}\int H'^T(\mu,\eta)e^{i\mu r}d\mu}"]
\end{tikzcd}$$

Fig. 3

The h' Circle of Transforms

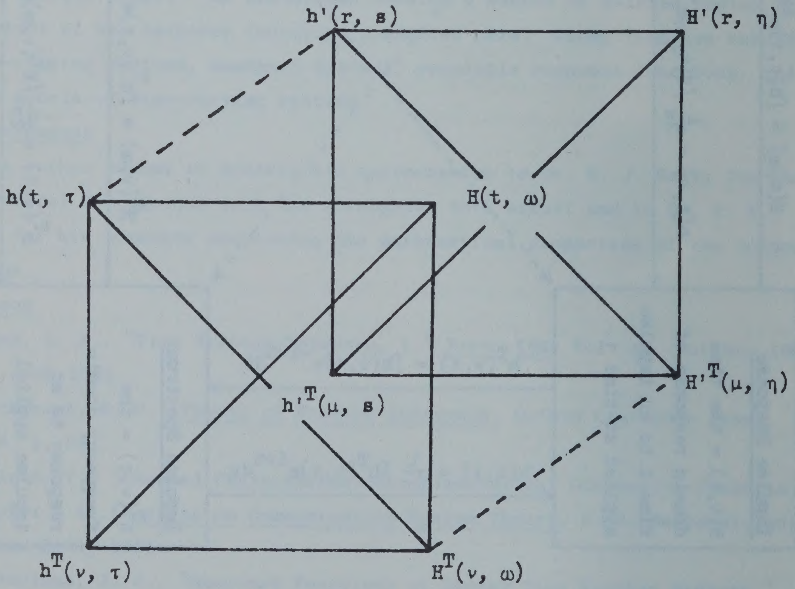


Fig. 4.

Relationships of the Primed Transforms
to the Unprimed Transforms

MATRIX REPRESENTATION OF LINEAR TIME-VARYING SYSTEMS

D. C. LAI
Department of Electrical Engineering
Northeastern University
Boston, Massachusetts

Summary - As has been observed in the analysis of linear time-varying systems, the basic difficulty seems to be a computational one rather than a theoretical one, and the lengthy computation is unavoidable. Therefore, it is desirable to concentrate the computational effort on obtaining particular solutions for different inputs as has been used in the classical methods of analysis. This paper describes an approximate representation of linear time-varying systems in matrix forms. This representation has the following advantages: (1) Once the matrix representing the system is obtained, the evaluation of the response to an input signal reduces to simple matrix multiplication. (2) It is easily adaptable to machine computation. (3) The matrix elements are directly measurable if the filters for generating basis functions are available. The properties of the matrix representation have been investigated. It is shown that the approximate representation is good for stable systems, and input signals which are of finite energy or sample functions of bounded random process. The accuracy of representation can be made to any desirable degree (in the least mean-square sense). Two simple examples are carried out for illustrative purpose.

I INTRODUCTION

In many engineering analyses time-varying systems, such as modulators, networks with time-varying elements, and certain control systems are encountered. They all have in common the property that the variation of the parameters of the system is caused by external effects, rather than by the inputs or dependent variables. The variation of the parameters may thus be described as functions of time or of some independent variables. The superposition law is still applicable for these systems. They are usually described by linear differential equations with time-varying coefficients, where time is considered as the independent variable. The response of the system may be obtained by solving the differential equations by well-known methods described in books on ordinary differential equations. These methods, however, are restricted to systems of low order or to certain special cases. The solution thus found applies only to the particular form of the input functions for which was obtained. For different input functions a completely new solution must be attempted.

This article describes an approximate representation of linear time-

This work was sponsored by the AFCRL, under Contract AF19(604)-7494, and by the AFOSR under Grant AF-AFOSR-62-371.

varying systems in matrix forms. Huggins[1] and Lai[2, 3] have discussed the feasibility of using this representation for time-varying systems and have applied this form to linear fixed systems. There are several advantages associated with this approximate representation: (i) Once the matrix representing the system is obtained,the evaluation of the response to an input signal reduces to simple matrix multiplication. (ii) It is easily adaptable to machine computation. (iii) The matrix elements are directly measurable if the filters for generating basis functions are available.

In a way, the proposed method is similar to that of Cruz and Van Valkenberg[4], Gilbert[5], Weiss[6]. They approximate the impulse response function (or weighting pattern) $h(t,\tau)$ of the system by separable system functions; i.e.,

$$H(t,\tau) = \sum_{i=1} C_i(t)\varphi_i(\tau) \tag{1.1}$$

where $\varphi_i(t)$ are a set of orthonormal functions and

$$C_i(t) = \int h(t,\tau)\varphi_i(\tau)\, d\tau . \tag{1.2}$$

On the other hand, if $h(t,\tau)$ is a Pincherle-Goursat kernel, it can be expressed exactly as the finite sum of the product of two separable functions. This class of systems has been treated in detail by Weiss[6]. The method is then essentially an approximation of the given $h(t,\tau)$ by the Pincherle-Goursat kernal.

The proposed method differs from the previous one in that it represents $h(t,\tau)$ in matrix form. The resulting output or the new system resulting by cascading two or more systems can be obtained through simple matrix multiplication. One more advantage is that the output, being in a decomposed form based on the same basis as was used for the input, can be used for further analysis.

II BACKGROUND

The linear time-varying systems considered here are describable by linear differential equations of the form

$$L(p,t)\, g(t) = M(p,t)\, f(t) \tag{2.1}$$

defined over an infinite or semi-infinite interval with prescribed initial conditions, where

$g(t)$ = system output,
$f(t)$ = input signal,
$L(p,t) = a_n(t)\, p^n + a_{n-1}(t)\, p^{n-1} + \ldots + a_o(t)$,
$M(p,t) = b_m(t)\, p^m + b_{m-1}(t)\, p^{m-1} + \ldots + b_o(t)$,
$p = \frac{d}{dt}$.

In the analysis problem most of the time we are interested in obtaining the response of the system to a given input signal when its state at some instant $t = t_o$ is specified. The state of a system is usually given by specifying the values of the output and its first (n-1) derivatives (if the system is of n^{th} order) at $t = t_o$.

If the system is assumed to be at rest initially, then the response is solely due to the input signal f(t) and may be expressed as

$$g(t) = \int_{-\infty}^{\infty} h(t,\tau)\, f(\tau)\, d\tau \qquad (2.2)$$

where $h(t,\tau)$ is the weighting function (or weighting pattern) of the time-varying system. It can be easily shown that $h(t,\tau)$ satisfies the equation

$$L(p,t)\, h(t,\tau) = M(p,t)\, \delta(t-\tau), \qquad (2.3)$$

i.e., $h(t,\tau)$ is the response of the system when a unit impulse is applied at $t = \tau$. The determination of this impulse response $h(t,\tau)$ from its differential equation has been discussed elsewhere[5, 7, 8]. The form of $h(t,\tau)$ thus obtained generally renders the integration (2.2) analytically insuperable.

Zadeh[9] has proposed a transform method which is based on the extension of the Fourier transform to time-varying systems. The transform of the time-domain system function is defined as

$$H(j\omega, t) = \int_{-\infty}^{\infty} h(t,\tau)\, e^{-j\omega(t-\tau)}\, d\tau, \qquad (2.4)$$

and then the output due to the input signal f(t) is obtained by

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(j\omega, t)\, F(j\omega)\, e^{j\omega t}\, d\omega\,. \qquad (2.5)$$

Unfortunately, this method shifts the computational difficulties to the determination of $H(j\omega, t)$ from $h(t,\tau)$ and also g(t) from $H(j\omega, t)$.

Gilbert[5] has proposed an approximation to $h(t,\tau)$ by separable functions as mentioned in the previous section. There are computational difficulties involved in obtaining the approximation, and the need of elaborate integration for finding the response to an input signal after the approximate form is obtained.

The basic difficulty in the analysis of linear time-varying systems is thus seen as a computational one rather than a theoretical one. It is also seen that lengthy computation is unavoidable. Then why should we not direct our computation effort to obtain results of a general nature. That is to say, why not obtain a representation of $h(t,\tau)$ which renders the integration (2.2) to a simple calculation good for a large class of input signals, rather than to obtain particular solutions of (2.1) for different input functions. These representations are generally approximations of $h(t,\tau)$, whose accuracy can be made to satisfy engineering applications. As has been mentioned previously, it is largely to this effort that the paper is directed. The representation proposed here will replace the integration (2.2) by a simple matrix multiplication.

III SYSTEM FUNCTION AS A MATRIX

Without loss of generality, we may assume that the system impulse response $h(t,\tau)$ is square-integrable and non-anticipatory; i.e.,

$$\int_{-\infty}^{\infty} h^2(t,\tau)\, d\tau < \infty \quad \text{and} \quad \int_{-\infty}^{\infty} h^2(t,\tau)\, dt < \infty,$$

and $h(t,\tau) = 0$ for $t < \tau$. Since this is true for stable physical systems, these conditions are satisfied for most practical problems.

We shall also assume that the input signals are physical signals and choose the time at which a signal starts as the origin; i.e.,

$$\int_{\infty}^{\infty} f^2(t)\, dt < \infty \text{ (finite energy)},$$

and $f(t) =$ for $t < 0$.
Then equation (2.2) becomes

$$g(t) = \int_0^{\infty} h(t,\tau)\, d\tau \quad \text{or} \quad g(t) = \int_0^{t} h(t,\tau)\, f(\tau)\, d\tau\,. \tag{3.1}$$

Representing $f(t)$ by

$$f(t) = \sum_{k=1}^{\infty} f_k \varphi_k(t). \tag{3.2}$$

which converges uniformly since $f(t)$ is of finite energy, and substituting (3.2) into (3.1), we obtain

$$g(t) = \int_0^{\infty} h(t,\tau) \sum_k f_k \varphi_k(\tau)\, d\tau = \sum_k f_k \int_0^{\infty} h(t,\tau)\varphi_k(\tau)\, d \tag{3.3}$$

where $f_k = \int_0^{\infty} f(t)\varphi_k(t)\, dt$ and $\varphi_k(t)$ is a member of an orthonormal set $\{\varphi\}$ complete in L_2-space. The projection of $g(t)$ on the j^{th} coordinate is then

$$g_j = \int_0^{\infty} g(t)\varphi_j(t)\, dt = \sum_k f_k \int_0^{\infty}\int_0^{\infty} \varphi_j(t)\, h(t,\tau)\varphi_k(\tau)\, dt\, d\tau\,. \tag{3.4}$$

With a slight change of notation, (3.4) becomes

$$g_j = \sum_k h_{jk} f_k$$

where

$$= \int_0^{\infty}\int_0^{\infty} \varphi_j(t)\, h(t,\tau)\varphi_k(\tau)\, dt\, d\tau \tag{3.5}$$

which may be recognized as the Hilbert quadratic form.

If now we consider $f(t)$ and $g(t)$ to be vectors in Hilbert space, and denote the collection of f_k's as a column vector which is the representation of the vector f(t) with respect to the orthonormal coordinate system $\{\varphi\}$, then the collection g_j's must also be a column vector with respect to $\{\varphi\}$. It is then obvious that the collection of h_{jk}, arranged in matrix form, must be the representation of the operator $h(t,\tau)$ with respect to $\{\varphi\}$, and

$$\begin{bmatrix} g_1 \\ g_2 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \cdots \\ h_{21} & h_{22} \cdots \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \quad \text{or} \quad |\underset{\sim}{g}> = \underset{\sim}{H}|\underset{\sim}{f}> \tag{3.6}$$

or

$$g(t) = \sum_j g_j\varphi_j(t) = \sum_j \sum_k h_{jk} f_k \varphi_j(t). \tag{3.7}$$

For practical purpose, the matrix $\underset{\sim}{H}$ and the vector $|\underset{\sim}{f}>$ have to be truncated.

IV PROPERTIES OF THE REPRESENTATION

Suppose $h(t,\tau)$ is symmetric and square-integrable as before, then $h(t,\tau)$ will have a set of eigenvalues λ_1, λ_2, ... associated with a set of orthonormal eigenfunctions ϕ_1, ϕ_2, ... If we use this eigenfunction set $\{\phi\}$ as the coordinate system, then by (3.5) we obtain

$$h_{jk} = \lambda_j \delta_{jk}$$

thus the $\underset{\sim}{H}$ matrix becomes diagonal. The component g_j of g(t) is expressed by

$$g_j = \lambda_j f_j \tag{4.1}$$

which implies that every component of g(t) coincides with every component of f(t) with a scalar change. The truncation of the series (3.2) has no effect on the component g_j. The response g(t) is then expressed by

$$g(t) = \sum_j \lambda_j f_j \phi_j(t). \tag{4.2}$$

It is seen that the accuracy of this expression depends solely on the number of terms in the series. This fact has been well expressed in the Hilbert-Schmidt Theorem that the series (4.2) converges absolutely and uniformly for every g(t) of type (3.1). The proof of this theorem is shown elsewhere.[10]

It can also be proved that any symmetric, square-integrable $h(t,\tau)$ either has an infinite number of eigenvalues or is expressible exactly as the finite sum of separable functions; i.e.,

$$h(t,\tau) = \sum_{k=1}^{n} \lambda_k \phi_k(t)\, \phi_k(\tau). \tag{4.3}$$

For the latter case, the $\underset{\sim}{H}$ matrix with respect to the orthonormal eigenset $\{\phi\}$ is

$$\underset{\sim}{H} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}, \tag{4.4}$$

which is a finite matrix.

For any arbitrary square-integrable system function $h(t,\tau)$, we can prove that

$$h(t,\tau) = \sum_{j=1}^{n} \sum_{k=1}^{n} h_{jk}\varphi_j(t)\varphi_k(\tau) \tag{4.5}$$

approximate $h(t,\tau)$ with any desired accuracy for large enough n; i.e., the mean-square error of the representation

$$E = \int_0^\infty\int_0^\infty \Big[h(t,\tau) - \sum_{j=1}^{n} \sum_{k=1}^{n} h_{jk}\varphi_j(t)\varphi_k(\tau)\Big]^2 \, dt \, d\tau$$

$$= \|h\|^2 - \|H\|^2 \tag{4.6}$$

can be kept as small as possible for sufficiently large n. The expression

$$\|H\|^2 = \sum_{j=1}^{n} \sum_{k=1}^{n} h_{jk}^2 \tag{4.7}$$

is usually defined as the square of the norm of a matrix.[12] We may state this formally as a theorem.

Theorem: Any square integrable system function $h(t,\tau)$; i.e., any system for which both

$$E_1^2(t) = \int_0^\infty h^2(t,\tau)\, d\tau, \qquad E_2^2(\tau) = \int_0^\infty h^2(t,\tau)\, dt$$

exist almost everywhere in the semi-infinite interval $(0,\infty)$ and are square integrable, can be decomposed (in an infinite number of ways) into the sum of

$$W(t,\tau) = \sum_{j=1}^{m} \sum_{k=1}^{n} h_{jk}\varphi_j(t)\varphi_k(\tau) \tag{4.8}$$

and another square integrable function $r(t,\tau)$ whose norm can be made smaller than any prescribed positive number ϵ; i.e.,

$$\|r\|^2 = \int_0^\infty\int_0^\infty r^2(t,\tau)\, dt\, d\tau < \epsilon^2. \tag{4.9}$$

Proof: Choose any orthonormal set $\{\varphi\}$ complete in the semi-infinite interval $(0,\infty)$. Let

$$\int_0^\infty h(t,\tau)\varphi_k(\tau)\, d\tau = y_k(t). \tag{4.10}$$

By Parseval's relation, we have

$$\sum_{k=1}^{\infty} y_k{}^2(t) = \int_0^{\infty} h^2(t,\tau)\, d\tau = E_1{}^2(t)$$

and consequently

$$\int_0^{\infty} y_k{}^2(t)\, dt \le \int \left[\sum_{k=1}^{\infty} y_k{}^2(t) \right] dt = \int E_1{}^2(t)\, dt = \int\!\!\int h^2(t,\tau)\, dt\, d\tau = \|h\|^2 \qquad (4.11)$$

which implies that $y_k(t)$ are square integrable. Hence, $y_k(t)$ can be represented by

$$y_k(t) = \sum_{j=1}^{\infty} C_j \varphi_j(t), \qquad \forall\ k$$

if we use the same complete set $\{\varphi\}$ as the basis functions to represent $y_k(t)$, where

$$C_j = \int_0^{\infty} y_k(t)\, \varphi_j(t)\, dt = \int_0^{\infty}\!\!\int_0^{\infty} \varphi_j(t)\, h(t,\tau)\, \varphi_k(\tau)\, dt\, d\tau = h_{jk}.$$

By virtue of the completeness of the set $\{\varphi\}$, the mean square error can be expressed by

$$\int_0^{\infty} \left[h(t,\tau) - \sum_{k=1}^{n} y_k(t)\, \varphi_k(\tau) \right]^2 d\tau = \sum_{k=n+1}^{\infty} y_k{}^2(t).$$

Integrate the above expression with respect to t, we have

$$\int_0^{\infty}\!\!\int_0^{\infty} \left[h(t,\tau) - \sum_{k=1}^{n} y_k(t)\, \varphi_k(\tau) \right]^2 dt\, d\tau = \sum_{k=n+1}^{\infty} \int_0^{\infty} y_k{}^2(t)\, dt$$

or

$$\int_0^{\infty}\!\!\int_0^{\infty} \left[h(t,\tau) - \sum_{j=1}^{m} \sum_{k=1}^{n} h_{jk}\, \varphi_j(t)\, \varphi_k(\tau) \right]^2 dt\, d\tau = \sum_{j=1}^{\infty} \sum_{k=n+1}^{\infty} h_{jk}{}^2 + \sum_{j=m+1}^{\infty} \sum_{k=1}^{n} h_{jk}{}^2. \qquad (4.12)$$

If we let

$$\int_0^{\infty} \varphi_j(t)\, h(t,\tau)\, dt = x_j(\tau), \text{ then}$$

$$\sum_{j=1}^{\infty} x_j{}^2(\tau) = \int_0^{\infty} h^2(t,\tau)\, dt = E_2{}^2(\tau)$$

and

$$\int_0^{\infty} x_j{}^2(\tau)\, d\tau \le \int_0^{\infty} \left[\sum_{j=1} x_j{}^2(\tau) \right] d\tau = \int_0^{\infty} E_2{}^2(\tau)\, d\tau = \int_0^{\infty}\!\!\int_0^{\infty} h^2(t,\tau)\, dt\, d\tau = \|h\|^2. \qquad (4.13)$$

Expressing $x_j(\tau) = \sum_{k=1}^{\infty} C_k \varphi_k(\tau)$, we have

$$C_k = \int_0^{\infty} x_j(\tau)\, \varphi_k(\tau)\, d\tau = \int_0^{\infty}\!\!\int_0^{\infty} \varphi_j(t)\, h(t,\tau)\, \varphi_k(\tau)\, dt\, d\tau = h_{jk}.$$

By the same token, we obtain

$$\int_0^\infty\int_0^\infty \Big[h(t,\tau) - \sum_{j=1}^{m} x_j(\tau)\,\varphi_j(t)\Big]^2 dt\, d\tau = \sum_{j=m+1}^{\infty} \int_0^\infty x_j^2(\tau)\, d\tau$$

or

$$\int_0^\infty\int_0^\infty \Big[h(t,\tau) - \sum_{j=1}^{m}\sum_{k=1}^{n} h_{jk}\varphi_j(t)\varphi_k(\tau)\Big]^2 dt\, d\tau = \sum_{j=m+1}^{\infty}\sum_{k=1}^{\infty} h_{jk}^2 + \sum_{j=1}^{m}\sum_{k=n+1}^{\infty} h_{jk}^2. \tag{4.14}$$

Put

$$\sum_{j=1}^{m}\sum_{k=1}^{n} h_{jk}\varphi_j(t)\varphi_k(\tau) = W(t,\tau)$$

and

$$h(t,\tau) - W(t,\tau) = r(t,\tau)$$

also

$$\sum_{j=1}^{\infty}\sum_{k=n+1}^{\infty} h_{jk}^2 + \sum_{j=m+1}^{\infty}\sum_{k=1}^{n} h_{jk}^2 = R_{mn} = \sum_{j=m+1}^{\infty}\sum_{k=1}^{\infty} h_{jk}^2 + \sum_{j=1}^{m}\sum_{k=n+1}^{\infty} h_{jk}^2$$

then

$$\|r\|^2 = \int_0^\infty\int_0^\infty r^2(t,\tau)\, dt\, d\tau = R_{mn}.$$

Since

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} h_{jk}^2 = \sum_{j=1}\int_0^\infty x_j^2(\tau)\, d\tau = \int_0^\infty E_2^2(\tau)\, d\tau = \|h\|^2$$

and

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} h_{jk}^2 = \sum_{k=1}\int_0^\infty y_k^2(t)\, dt = \int_0^\infty E_1^2(t)\, dt = \|h\|^2,$$

R_{mn} can be made as small as possible for large enough n. Therefore, R_{mn} can be made smaller than ϵ^2 for sufficiently large m and n.

For practical purposes, we shall let m = n, then the mean-square error E can be made as small as possible. If the input signals f(t) are in L_2, then f(t) can be represented by the series

$$f(t) = \sum_{k=1}^{n} f_k\varphi_k(t)$$

with sufficient accuracy for large n, and therefore the output g(t) represented by

$$g(t) = \sum_{j=1}^{n}\sum_{k=1}^{n} h_{jk} f_k \varphi_j(t)$$

will be sufficiently accurate. In most of the practical cases, the input signals are of finite energy and the systems are stable. We see then that the representation is good for a large class of practical signals.

The expression (4.6) does, at least, give a measure as to how close the approximation matrix is to $h(t,\tau)$ for a particular n. Unfortunately, there is no explicit expression for the approximation error

$$\int_0^\infty \left[g(t) - \sum_{j=1}^{n} \sum_{k=1}^{n} h_{jk} f_k \varphi_j(t)\right]^2 dt$$

in terms of the approximation error of $h(t,\tau)$.

To summarize, we have shown that the representation will be exact for symmetric physical realizable $h(t,\tau)$ with a finite number of eigenvalues. For any other stable systems, the representation gives sufficient accuracy and is good for a large class of input signals. The assertion made at the beginning is thus justified.

Since there are many ways to represent $h(t,\tau)$ and the input $f(t)$, it is reasonable to ask which orthonormal set complete in L_2-space will give the most efficient representation for both $h(t,\tau)$ and $f(t)$, and thus for the response $g(t)$; i.e., which set will yield a prescribed mean-square error with the least number n? It is also legitimate to ask how can one find such an orthonormal set. These questions will need further investigation.

V REPRESENTATION BY TWO SETS OF COORDINATE SYSTEMS

It is interesting to note that in (3.5),

$$h_{jk} = \int_0^\infty \int_0^\infty \varphi_j(t)\, h(t,\tau) \varphi_k(\tau)\, dt\, d\tau,$$

φ_j and φ_k do not necessarily belong to the same set. That is, we may use the bilinear form as well as the quadratic form. To distinguish them, we shall denote one set by $\{\psi\}$ and the other by $\{\varphi\}$, then we have

$$h_{jk} = \int_0^\infty \int_0^\infty \psi_j(t)\, h(t,\tau) \varphi_k(\tau)\, dt\, d\tau . \qquad (5.1)$$

For practical reasons, we may choose a set, say $\{\varphi\}$, to represent the input $f(t)$ efficiently and another set $\{\psi\}$ to represent $y_k(t) = \int_0^\infty h(t,\tau)\, \psi_k(\tau)\, d\tau$ efficiently. The most convenient and efficient set for $y_k(t)$ would be the orthonormalized set constructed from the $y_k(t)$'s if they are linearly independent. The Gram-Schmidt process may be used for the orthonormalization of the $y_k(t)$. However, the computation will be much more complicated.

If $y_k(t)$ are linearly dependent, the set can be made linearly independent by discarding the unnecessary elements. These discarded elements can always be expressed as the linear combination of the elements of the orthonormalized $\{\psi\}$.

In this way, the available $y_k(t)$ can be expressed exactly by the orthonormal set $\{\psi\}$. The approximation is thus solely due to our confining to the subspace spanned by the truncated $\{\varphi\}$ set for practical reasons.

VI SYNTHESIS OF THE SYSTEM FROM ITS MATRIX

The projection of the input f(t) on the coordinate system $\{\varphi\}$ may be obtained in a simple way. The projection of f(t) onto the kth coordinate is expressed by

$$f_k = \int_0^\infty f(t)\,\varphi_k(t)\,dt \qquad (6.1)$$

The output of a filter with impulse response $\varphi_k(t)$ for the input f(-t) is

$$v_k(t) = \int_0^\infty f(\tau - t)\,\varphi_k(\tau)\,d\tau\,.$$

If we evaluate $v_k(t)$ at the instant t = 0, we obtain

$$v_k(0) = \int_0^\infty f(\tau)\,\varphi_k(\tau)\,d\tau = f_k,$$

which is a well-known result. A way of implementing this has been carried out by Huggins[1]: Record f(t) on a tape and play it back through the filter φ_k with the tape reversed and then sample the output of the filter at t = 0. Use of this method can greatly simplify the simulation of the system.

We may synthesize a given system in matrix form by either an analog or a digital computer. For the same input, the simulated system will give an output which is different from that of the actual system by a specified square-mean error. The diagram in Fig. 1 depicts a scheme for the simulation of $h(t,\tau)$ when n = 5.

VII RESPONSE TO RANDOM INPUT FUNCTIONS

At this point, we shall examine the output when the input is a sample function of some random process. We shall assume that the generalized convolution integral

$$g(t) = \int_0^\infty h(t,\tau)\,f(\tau)\,d\tau$$

exists for every sample function f(t) in the given random process. It is necessary that the given random process is bounded and the system is stable. It follows then that the approximate representation

$$g(t) = \sum_j g_j\,\varphi_j(t), \qquad (7.1)$$

where

$$g_j = \sum_k h_{jk}\,f_k, \qquad (7.2)$$

is valid for the input sample function f(t). Of course, g_j and f_j are random coefficients. For convenience, we shall assume all the g_j, f_j and $\varphi_j(t)$ are real as before. (The analysis can be easily generalized to complex numbers and complex functions.) In most practical cases, we are interested in the correlation functions involving the input and the output. These can be evaluated by applying the matrix representation of the system. Suppose the output autocorrelation function is desired. We shall show that it can be evaluated in a simple manner. By definition, the autocorrelation is

$$h_g(t, q) = E(g_t\, g_q). \tag{7.3}$$

By virture of (7.1),

$$R_g(t, q) = E\left[\sum_j g_j \varphi_j(t) \sum_k g_k \varphi_k(q)\right]$$

$$= \sum_j \sum_k \varphi_j(t)\, \varphi_k(q)\, E(g_j\, g_k). \tag{7.4}$$

Using (7.2), we obtain

$$E(g_j\, g_k) = E\left[\sum_\ell h_{j\ell} f_\ell \sum_m h_{km} f_m\right]$$

$$= \sum_\ell \sum_m h_{j\ell}\, h_{km}\, E(f_\ell\, f_m). \tag{7.5}$$

Let the autocorrelation function of the input random process be denoted by $R_f(t, q)$ which is known. Then we may write

$$R_f(t, q) = E(f_t f_q) = E\left[\sum_\ell f_\ell \varphi_\ell(t) \sum_m f_m \varphi_m(q)\right]$$

$$= \sum_\ell \sum_m \varphi_\ell(t)\, \varphi_m(q)\, E(f_\ell\, f_m). \tag{7.6}$$

Representing $R_f(t, q)$ in matrix form with respect to the orthonormal set $\{\varphi\}$, we obtain its element

$$R^f_{\ell m} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_\ell(t)\, R_f(t, q) \varphi_m(q)\; dt\; dq. \tag{7.7}$$

From (7.6) it follows that

$$R^f{}_m = E(f\;\, f_m). \tag{7.8}$$

Combining (7.4), (7.5) and (7.8), we obtain

$$R_g(t, q) = \sum_{j, k, \ell, m} \varphi_j(t) \varphi_k(q)\; h_{j\ell}\, h_{km}\; R^f_{\ell m} \tag{7.9}$$

which is equivalent to

$$R_g(t, q) = \left[\varphi_1(t), \ldots\right] \begin{bmatrix} h_{11} & h_{12} & \ldots \\ h_{21} & h_{22} & \ldots \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{bmatrix} \begin{bmatrix} R_{11}{}^f & R_{12}{}^f & \ldots \\ R_{21}{}^f & R_{22}{}^f & \ldots \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{bmatrix} \begin{bmatrix} h_{11} & h_{21} & \ldots \\ h_{12} & h_{22} & \ldots \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{bmatrix} \begin{bmatrix} \varphi_1(q) \\ \varphi_2(q) \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

$$= \langle \varphi(t) \,|\, \underset{\sim}{H} \,|\, \underset{\sim}{R}^f \,|\, \underset{\sim}{H}^t \,|\, \varphi(q) \rangle . \tag{7.10}$$

Hence the matrix representation of the output autocorrelation is

$$\underset{\sim}{R}^g = \underset{\sim}{H} \,|\, \underset{\sim}{R}^f \,|\, \underset{\sim}{H}^t. \tag{7.11}$$

It is interesting to note that we would obtain the Karhunen-Loéve expansion of the sample function of the input if the orthonormal set $\{\varphi\}$ chosen were the eigenset of $R_f(t, q)$.[11]

VIII DETERMINATION OF THE ELEMENTS OF THE MATRIX

In general, the system to be analyzed is either described by the differential equation (2.1), or it is in a black box with input and output leads accessible; very seldom is the impulse response $h(t,\tau)$ known. For many cases, even the differential equation is hard to obtain. Furthermore, it is rather rare that an analytic expression for $h(t,\tau)$ can be obtained. Machine computations are necessary for obtaining $h(t,\tau)$ from equation (2.3) and for the evaluation of (3.5) for h_{jk}. It can be shown that the function $\int h(t,\tau)\,\varphi_k(\tau)\,d\tau$ is periodic if the differential operators L and M have periodic coefficients. Then for the evaluation of h_{jk} only the fundamental period has to be considered. The procedure would be: (1) machine solution of equation (2.1), and (2) machine evaluation of expression (3.5). Usually, the computation is routine and we shall not elaborate on it here.

If the system is in a physical form with input and output leads accessible, direct measurement of the matrix elements h_{jk} is possible provided the orthonormal filters for the basis functions are available. This method is particularly suitable if simulation of the system according to the differential equation on an analog computer is possible. We have assumed for this analysis that the system is in its ground state. It is, therefore, necessary for the system to return to the ground state after each h_{jk} has been measured. The process is described in Fig. 2. This is made possible since by (3.5)

$$h_{jk} = \int_0^\infty\int_0^\infty \varphi_j(t)\, h(t,\tau)\, dt\, d\tau$$

$$= \int_0^\infty \varphi_j(t)\, y_k(t)\, dt,$$

which is equivalent to

$$\int_0^\infty \varphi_j(\tau)\, y_k(\tau - t)\, d\tau \quad \text{at } t = 0.$$

We see now that the choice of a coordinate system in which the system is going to be represented depends on many factors. Those which we wish to consider are: (i) the number of approximation terms (dimensionality of subspace) for a specified accuracy; (ii) the work involved in evaluating the integration (3.3); (iii) the effort involved in the evaluation of (3.5); and (iv) the availability of orthonormal filters for the particular waveforms. We would, of course, like to minimize (i), (ii) and (iii), and have (iv) available either in digital or analog form. However, any set which satisfies one of these conditions does not necessarily satisfy all the others. For instance, the set which will minimize the number of terms is the set that differs the least from the eigenset of the system in consideration; this set guarantees neither simplicity of evaluating (3.3) or (3.5), nor the availability in (iv). Unfortunately, little can be said about the general relation between (i), (ii), (iii) and (iv).

IX ILLUSTRATIVE EXAMPLES

Two simple examples are shown below. The first one is a simple RC network with a time-varying resistor. The second one is a second-order control system which has been treated by Farmanfarma[13] and Gilbert[5].

Example 1: A simple RC circuit, shown in Fig. 3(a) has a time-varying resistor r(t). The differential equation between the input and output is found to be

$$h(t,\tau) = \delta(t-\tau) - \frac{g(\tau)}{C}\, e^{\frac{1}{C}\int_t^\tau g(\eta)\,d\eta} \quad , t > \tau$$

$$= 0 \quad , t < \tau$$

where $g(t) = \frac{1}{r(t)}$.

For simplicity, in the following computation for the matrix of $h(t,\tau)$ we shall let C = 1 and g(t) have the form shown in Fig. 4. The circuit becomes that of Fig. 3(b). The basis used here is a set of orthonormalized exponentials consisting

$$\varphi_1(t) = \sqrt{2}\, e^{-t}$$
$$\varphi_2(t) = 2(-2e^{-t} + 3e^{-2t} + 10e^{-3t})$$
$$\varphi_3(t) = \sqrt{6}\,(3e^{-t} - 12e^{-2t} + 10e^{-3t})$$
$$\varphi_4(t) = 2\sqrt{2}\,(-4e^{-t} + 30e^{-2t} - 60e^{-3t} + 35e^{-4t})$$
$$\varphi_5(t) = \sqrt{10}\,(5e^{-t} - 60e^{-2t} + 210e^{-3t} - 280e^{-4t} + 126e^{-5t}).$$

The matrix obtained is

$$H = \begin{bmatrix} .076905 & -.100244 & .072744 & .152947 & -.026849 \\ .389073 & .356150 & -.093513 & -.144530 & -.005306 \\ .146982 & .462385 & .351629 & -.117196 & .084087 \\ .019283 & -.004780 & .517863 & .666620 & -.127818 \\ .030128 & .067602 & .062611 & .419571 & .555666 \end{bmatrix}$$

Using $\varphi_1(t)$ as an input, we obtain the output vector

$$| V_o\rangle = \{0.076905 \quad 0.389073 \quad 0.146982 \quad 0.019283 \quad 0.030128\}$$

or $$v_o(t) = 0.076905\,\varphi_1(t) + 0.389073\,\varphi_2(t) + 0.146982\,\varphi_3(t) + 0.019283\,\varphi_4(t) + 0.030128\,\varphi_5(t).$$

The approximated $v_o(t)$ obtained by matrix multiplication and $v_o(t)$ obtained by integration are plotted in Fig. 5. Apparently, the basis functions have smoothed out the kinks in the exact response. The mean-square error of the response is

$$\int_0^\infty v_o^2(t)\,dt - \sum_{j=1}^{5} v_j^2 = 0.1865 - 0.1802 = 0.0063,$$

which is about 3.4% of the total energy.

Example 2: A second-order control system shown in Fig. 6 is used. The input-output is found to be related by the differential equation

$$\frac{d^2v}{dt^2} + 5\frac{dv}{dt} + g(t)\,v(t) = g(t)\,f(t).$$

The time-varying element g(t) has the same waveform as that shown in Fig. 4. By using the same set of basis functions as those used in Example 1, we obtain the matrix for $h(t,\tau)$ as follows:

$$H = \begin{bmatrix} .503452 & .114381 & .017881 & -.077248 & -.033817 \\ -.588044 & .103302 & -.003848 & .126408 & .074419 \\ .227481 & -.358628 & .151255 & -.040077 & -.058287 \\ -.076203 & .286715 & -.308259 & .049263 & .027751 \\ .071452 & -.165679 & .233854 & -.186594 & -.068349 \end{bmatrix}.$$

A triangular wave is used as an input whose vector form with respect to $\{\varphi\}$ is found to be

$$|\underset{\sim}{f}\rangle = \{1.067123 \quad -.393177 \quad .169153 \quad -.064707 \quad .010852\}.$$

The output obtained by matrix multiplication is

$$|\underset{\sim}{v}\rangle = \{.499930 \quad -.676154 \quad .411301 \quad -.249077 \quad .193762\}.$$

Both the input and the response are plotted in Fig. 7.

X CONCLUSIONS

In this article, an expository account of the properties and usefulness of matrix representation of linear time-varying systems have been given. It has also been shown that the approximate representation is good for stable systems, and input signals which are of finite energy or sample functions of bounded random process. The accuracy of representation can be made to any desirable degree (in the least mean-square sense). There remains the problem of finding an appropriate basis to satisfy a specified criterion.

The advantage of using the matrix representation has been demonstrated. It requires virtually no effort in evaluating the response of a linear time-varying system when the matrix for the system is obtained, though the determination of the matrix elements must be done by machine computation. It would be most advantageous for use in the case when the system is given in a physical form, since the matrix elements can be determined by direct measurements.

ACKNOWLEDGEMENT

The author wishes to thank Mr. Eric Reid for his patience in programming for the computation of the illustrative examples.

REFERENCES

1. W.H. Huggins, "Part I. The Use of Orthogonalized Exponentials", AD-133741 (November 15, 1958).
2. D.C. Lai, "Part VI. Signal-Space Concepts and Dirac's Notation", AD-235151 (January 15, 1960).
3. D.C. Lai, "Part II. An Orthonormal Filter for Exponential Waveforms", AD-152443, (June 15, 1958).
4. J.B. Cruz, Jr., and M.F. Van Valkenburg, "The Synthesis of Models for Time-Varying Linear Systems", Proc. Symp. on Active Networks and Feedback Systems, Poly. Inst. of Brooklyn, Brooklyn, New York, (April 1960).
5. E.G. Gilbert, .An Approximate Method for Analytically Evaluating the Response of Time-Varying Linear Systems", IRE Trans. on Circuit Theory vol. CT-9, No. 3, pp289-295, (September 1961).
6. L. Weiss, "Respresentation and Analysis of Signals, Part XIV, Time-Varying Systems with Separable System Functions", (report), The Johns Hopkins University, Department of Electrical Engineering, Baltimore, MD. (January 30, 1963).
7. L.A. Zadeh, "The Determination of the Impulsive Response of Variable Networks", Journal of Applied Physics, vol. 21, No. 7, pp. 642-645, (July 1950).
8. K.S. Miller, "Properties of Impulsive Responses and Green's Functions," IRE Trans. on CT, CT-2, pp. 26-32, (March 1955).
9. L.A. Zadeh, "Frequency Analysis of Variable Networks", Proc. IRE, vol. 38, No. 3, pp. 291-299, (March 1950).
10. F.G. Tricomi, "Integral Equations", Intersciences Publishers, New York, New York, pp. 110-118, (1957).
11. Davenport and Root, "An Introduction to the Theory of Random Signals and Noise, McGraw-Hill, New York, New York, pp. 96-99, (1958).
12. R. Bellman, Introduction to Matrix Analysis, McGraw-Hill, New York, New York, p. 162, (1960).
13. G. Farmanfarma, "General Analysis and Stability Study of Finite Pulsed Feedback Systems", Trans. AIEE, vol. 77, Application and Industry, pp. 148-162, (July 1958).

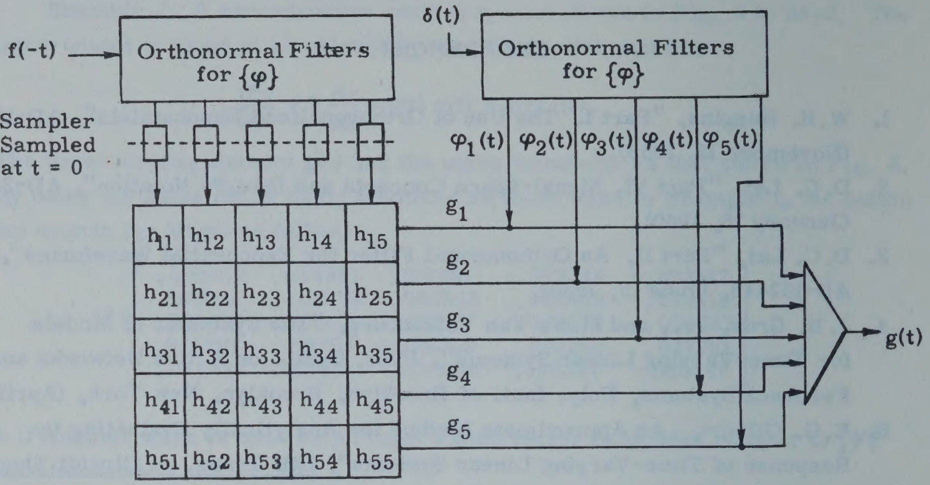

f(-t)
Orthonormal Filters for {φ}
δ(t)
Orthonormal Filters for {φ}
Sampler
Sampled
at t = 0
φ1(t)
φ2(t)
φ3(t)
φ4(t)
φ5(t)
h11 h12 h13 h14 h15
h21 h22 h23 h24 h25
h31 h32 h33 h34 h35
h41 h42 h43 h44 h45
h51 h52 h53 h54 h55
g1
g2
g3
g4
g5
g(t)

Fig. 1 Simulation of a time-varying system $h(t,\tau)$ whose matrix representation $\underset{\sim}{H}$ is given.

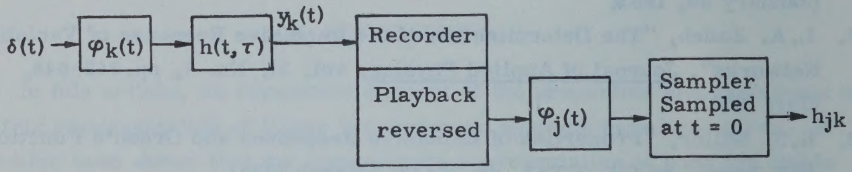

δ(t)
φk(t)
h(t, τ)
yk(t)
Recorder
Playback
reversed
φj(t)
Sampler
Sampled
at t = 0
hjk

Fig. 2 Direct measurement of the matrix element h_{jk}.

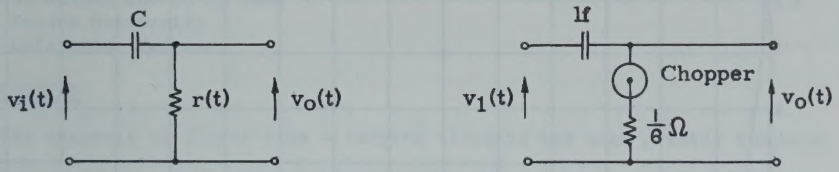

C
vi(t)
r(t)
vo(t)
1f
v1(t)
Chopper
1/6 Ω
vo(t)

Fig. 3 Simple RC networks with time-varying resistor.

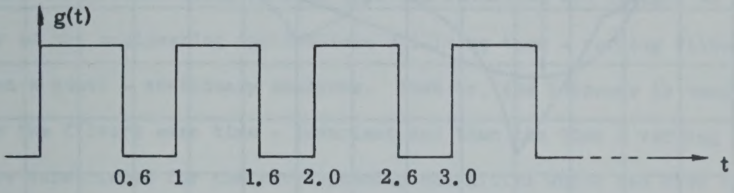


Fig. 4 Waveform used as the time-varying element.

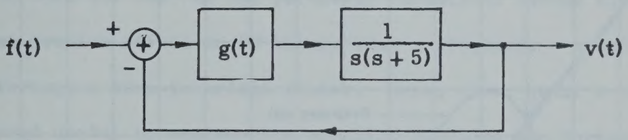

f(t)
+
−
g(t)
1/s(s + 5)
v(t)

Fig. 6 A second-order control system with finite-width sampling.

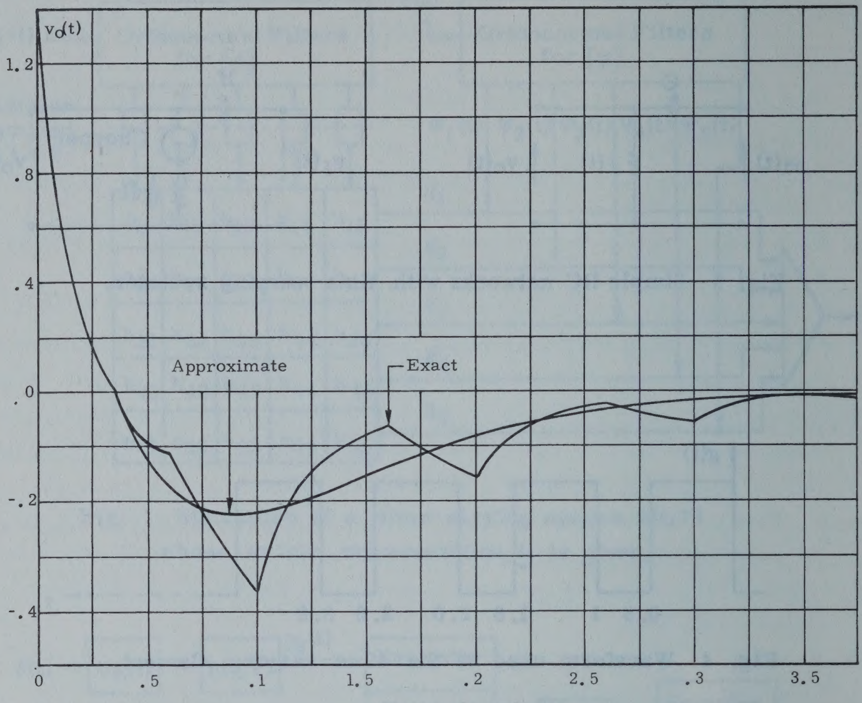


Fig. 5 Response of $h(t,\tau)$ to a decaying exponential $\sqrt{2}e^{-t}$.

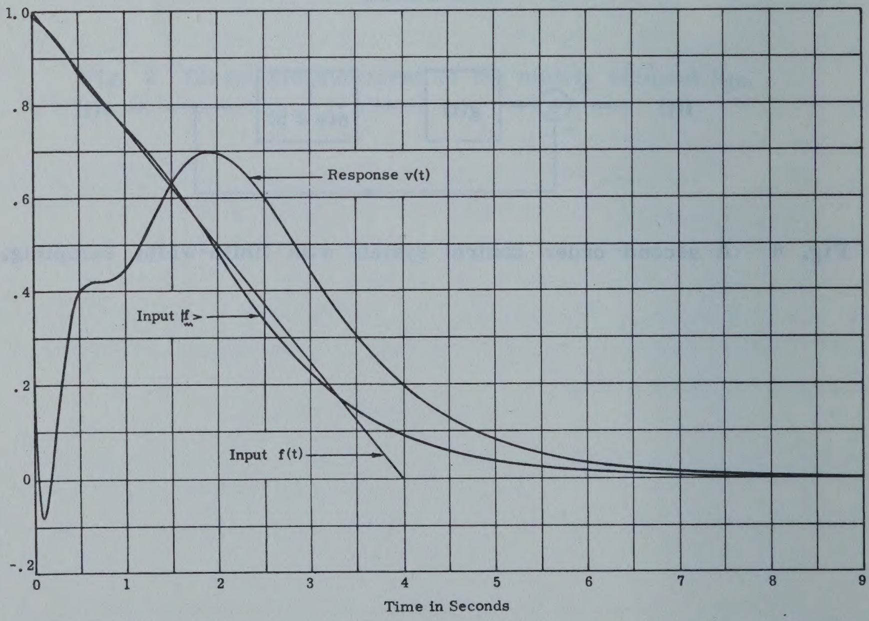


Fig. 7 Input and response of the second-order control system as obtained by matrix multiplication.

ANALYSIS OF TIME-VARYING SYSTEMS: AN ANOLOGY WITH WITH TIME-INVARIANT SYSTEMS

D. WEINER and B. J. LEON
Purdue University
Lafayette, Indiana

INTRODUCTION

The analysis of linear time - varying circuits has been greatly hampered by the fact that, except for the first -order case and some special forms of the second - order case, no closed form solutions are available for linear differential equations with variable coefficients. Because of the mathematical difficulties encountered in the solution of such equations, it appears desirable to approach the problem using as much physical intuition and insight as possible.

Many of the engineering applications utilizing time - varying filters are based upon a quasi - stationary analysis. That is, the response is computed as though the filters were time - invariant and then the time - varying parameters are substituted for the corresponding quantities which had been assumed constant. Heuristically, we suspect that this analysis is valid if the parameter variation is "slow enough" - but how slow is "slow enough"? This question can be answered provided there exists an exact analysis which yields a solution in terms of a quasi - stationary response function plus (or times) a correction term that goes to zero (or one) as the parameter variations become slow.

An analogous problem arises in FM theory when a frequency - modulated wave is passed through a time-invariant filter. Here, also, quasi - stationary analysis has proved useful in engineering practice provided the instantaneous frequency of the FM excitation varies "slowly enough." The quasi-stationary term is computed as if the FM signal had no modulation. Then the instantaneous frequency is inserted in the response function in place of the original fixed frequency. Concerning this FM problem, various authors [1-6] have tried to get useful results in the form of a quasi-stationary term plus a correction term. The correction term in references [1 - 4] is presented as an infinite series expansion. Reference [5] gives an approximate correction term in

---

*This Research was sponsored by the National Science Foundation under Grant GP-581.

closed form. A closed form expression is given in reference [6]; but, as presented, the result is incorrect.

This paper points out the similarity between the responses of time-invariant filters excited by frequency modulated wave forms and time-varying filters excited by sinusoids. In particular, an exact closed form solution is presented for a time-invariant second-order filter excited by an arbitrary FM signal. Also, for first-order time-varying filters, exact solutions are presented having the same form as that of the FM time-invariant filter problem. In each case the solutions are expressed as the sum of a quasi-stationary term plus a correction term. When the time-varying parameters change "slowly enough", the correction term becomes negligible. When the parameter variations occur at a more rapid rate, the correction term grows in magnitude. It then interferes with the quasi-stationary term to produce significant frequency and amplitude fluctuations in the over-all response. The method used to derive these closed form expressions begins with a piece-wise constant approximation of the time-varying parameters and refines the approximation so that the piece-wise constant solution equals the solution for the continuous case. All solutions presented have been verified by direct substitution into the appropriate differential equation.

RESPONSE OF A SINGLE-TUNED CIRCUIT TO A FREQUENCY-MODULATED SINUSOID

The first problem under consideration is shown in Figure 1. The quasi-stationary impedance $Z[j\omega_i(t)]$, is defined to be the conventional steady-state impedance in which the steady-state frequency variable, ω, is replaced by the time-varying instantaneous frequency of the excitation, $\omega_i(t)$. $\omega_i(t)$ is assumed to be constant from $t = -\infty$ until $t = o$. For $t > o$, $\omega_i(t)$ is allowed to vary in some arbitrary but specified manner. The step in frequency at $t = o$ occurs because the slope of the instantaneous phase angle, $\theta_i(t)$, is not required to be continuous at $t = 0$. More specifically,

$$\theta_i(t) = \begin{cases} \omega_\nu t & t \leq 0 \\ \int_o^t \omega(\xi)\, d\xi = \theta(t) & t > 0 \end{cases} \qquad (1)$$

$$\omega_i(t) = \frac{d\theta_i}{dt} = \begin{cases} \omega_\nu & t \le 0 \\ \omega(t) = \frac{d\theta}{dt} & t > 0 \end{cases} \tag{2}$$

The solution begins by making a piece-wise constant approximation of $\omega_i(t)$ as shown in Figure 2. Denote the approximation by $\hat{\omega}(t)$. During the time interval, $n\,\Delta\,\tau \le t < (n + 1)\,\Delta\,\tau$, the frequency of the piece-wise constant wave form is ω_n. The corresponding phase angle is $\hat{\theta}(t) = \sum_{u=0}^{n-1} \omega_u\,\Delta\,\tau + \omega_n(t - n\Delta\tau)$. The response in this interval consists of a particular solution plus a homogeneous solution of the corresponding second-order differential equation with constant coefficients. In particular,

$$e_o(t) = \underbrace{Z(j\omega_n)\, e^{j[\sum_{u=0}^{n-1} \omega_u \Delta\tau + \omega_n(t - n\Delta\tau)]}}_{\text{particular solution}} + \underbrace{K_1^n\, e^{(-\alpha + j\omega_d)(t - n\Delta\tau)} + K_2^n\, e^{(-\alpha - j\omega_d)(t - n\Delta\tau)}}_{\text{homogeneous solution}} \tag{3}$$

The two arbitrary constants are evaluated by noting that the voltage across the capacitor and the flux linkage through the inductor are continuous at $t = n\Delta\tau$. K_1^n and K_2^n are given by

$$K_1^n = \begin{cases} \left(1 + j\frac{\alpha}{\omega_d}\right)\left\{\frac{Z(j\omega_o)}{2}\left[-\left(\frac{\omega_d + \omega_o}{\omega_o}\right) + j\frac{\alpha}{\omega_o}\right] - \frac{Z(j\omega_\nu)}{2}\left[-\left(\frac{\omega_d + \omega_\nu}{\omega_\nu}\right) + j\frac{\alpha}{\omega_\nu}\right]\right\} & n=o \\ \left(1 + j\frac{\alpha}{\omega_d}\right)\left\{\frac{Z(j\omega_n)}{2}\left[-\left(\frac{\omega_d + \omega_n}{\omega_n}\right) + j\frac{\alpha}{\omega_n}\right] - \frac{Z(j\omega_{n-1})}{2}\left[-\left(\frac{\omega_d + \omega_{n-1}}{\omega_{n-1}}\right) + j\frac{\alpha}{\omega_{n-1}}\right]\right\} e^{j[\sum_{u=0}^{n-1} \omega_u \Delta\tau]} & \\ \quad + K_1^{n-1}\, e^{(-\alpha + j\omega_d)\Delta\tau} & n = 1, 2, 3, \ldots \end{cases} \tag{4}$$

$$K_2^n = \begin{cases} \left(1 - j\frac{\alpha}{\omega_d}\right)\left\{\frac{Z(j\omega_o)}{2}\left[\left(\frac{\omega_d - \omega_o}{\omega_o}\right) + j\frac{\alpha}{\omega_o}\right] - \frac{Z(j\omega_\nu)}{2}\left[\left(\frac{\omega_d - \omega_\nu}{\omega_\nu}\right) + j\frac{\alpha}{\omega_\nu}\right]\right\} & n=0 \\ \left(1 - j\frac{\alpha}{\omega_d}\right)\left\{\frac{Z(j\omega_n)}{2}\left[\left(\frac{\omega_d - \omega_n}{\omega_n}\right) + j\frac{\alpha}{\omega_n}\right] - \frac{Z(j\omega_{n-1})}{2}\left[\left(\frac{\omega_d - \omega_{n-1}}{\omega_{n-1}}\right) + j\frac{\alpha}{\omega_{n-1}}\right]\right\} e^{j[\sum_{u=0}^{n-1} \omega_u \Delta\tau]} & \\ \quad + K_2^{n-1}\, e^{(-\alpha - j\omega_d)\Delta\tau} & n = 1, 2, 3, \ldots \end{cases} \tag{5}$$

It should be emphasized that although Equation (3) is valid for all integer n greater than or equal to zero, once n is specified the response is valid only in the interval $n\Delta\tau \leq t < (n+1)\Delta\tau$. The expressions for K_1^n and K_2^n are simplified by defining the following quantities:

$$A = (1 + j\frac{\alpha}{\omega_d}) \qquad A^* = (1 - j\frac{\alpha}{\omega_d}) \qquad B_n = \frac{Z(j\omega_n)}{2}$$

$$C_n = \left[-\left(\frac{\omega_d + \omega_n}{\omega_n}\right) + j\frac{\alpha}{\omega_n}\right] \qquad D_n = \left[\left(\frac{\omega_d - \omega_n}{\omega_n}\right) + j\frac{\alpha}{\omega_n}\right]$$

$$E_n = e^{j[\sum_{u=0}^{n-1} \omega_u \Delta\tau]} \qquad F = e^{(-\alpha + j\omega_d)\Delta\tau} \qquad F^* = e^{(-\alpha - j\omega_d)\Delta\tau}$$

It follows that

$$K_1^n = \begin{cases} A[B_oC_o - B_\nu C_\nu] & n = 0 \\ A[B_nC_n - B_{n-1}C_{n-1}]\,E_n + K_1^{n-1}\,F & n = 1, 2, 3, \ldots \end{cases} \qquad (6)$$

$$K_2^n = \begin{cases} A^*[B_oD_o - B_\nu D_\nu] & n = 0 \\ A^*[B_nD_n - B_{n-1}D_{n-1}]\,E_n + K_2^{n-1}\,F & n = 1, 2, 3, \ldots \end{cases} \qquad (7)$$

Equations (6) and (7) are recursion relations in the unknown quantities K_1^n and K_2^n respectively. They can be combined to yield

$$K_1^n = \begin{cases} A[B_oC_o - B_\nu C_\nu] & n = 0 \\ AF^n[B_oC_o - B_\nu C_\nu] + \sum_{\ell=1}^{n} A[B_\ell C_\ell - B_{\ell-1}C_{\ell-1}](F)^{n-\ell}\,E_\ell & n = 1, 2, 3, \ldots \end{cases} \qquad (8)$$

$$K_2^n = \begin{cases} A^*[B_oD_o - B_\nu D_\nu] & n = 0 \\ A^*(F^*)^n\,[B_oD_o - B_\nu D_\nu] + \sum_{\ell=1}^{n} A^*\,[B_\ell D_\ell - B_{\ell-1}D_{\ell-1}](F^*)^{n-\ell}\,E_\ell & n = 1, 2, 3, \ldots \end{cases} \qquad (9)$$

The solution for the continuous case is obtained by refining the approximation in such a manner that $n\Delta\tau \to t$, $\ell\Delta\tau \to \tau$, and $u\Delta\tau \to \xi$. In this refinement the region of validity of Equation (3) shrinks from the interval $n\Delta\tau \leq t < (n+1)\Delta\tau$ to the single instant of time, t. However, since t is arbitrary, Equation (3) becomes valid for all $t \geq 0$. Specifically,

$$e_o(t) = \underbrace{Z[j\omega(t)]e^{j\int_0^t \omega(\xi)d\xi}}_{\text{quasi-stationary term}} + \underbrace{K_1(t) + K_2(t)}_{\text{correction term}} \qquad (10)$$

where

$$K_1(t) = \left(1 + j\frac{\alpha}{\omega_d}\right)\left\{\frac{Z(j\omega_o)}{2}\left[-\left(\frac{\omega_d+\omega_o}{\omega_o}\right) + j\frac{\alpha}{\omega_o}\right]\right.$$

$$\left.-\frac{Z(j\omega_\nu)}{2}\left[-\left(\frac{\omega_d+\omega_\nu}{\omega_\nu}\right) + j\frac{\alpha}{\omega_\nu}\right]\right\}e^{(-\alpha+j\omega_d)t} + \left(1 + j\frac{\alpha}{\omega_d}\right)\int_0^t \frac{d}{d\tau}\left\{\frac{Z[j\omega(\tau)]}{2}\right.$$

$$\left.\left[-\left(\frac{\omega_d+\omega(\tau)}{\omega(\tau)}\right) + j\frac{\alpha}{\omega(\tau)}\right]\right\}e^{(-\alpha+j\omega_d)(t-\tau)}\, e^{j\int_0^\tau \omega(\xi)d\xi}\, d\tau \qquad (11)$$

and

$$K_2(t) = \left(1 - j\frac{\alpha}{\omega_d}\right)\left\{\frac{Z(j\omega_o)}{2}\left[\left(\frac{\omega_d-\omega_o}{\omega_o}\right) + j\frac{\alpha}{\omega_o}\right]\right.$$

$$\left.-\frac{Z(j\omega\nu)}{2}\left[\left(\frac{\omega_d-\omega_\nu}{\omega_\nu}\right) + j\frac{\alpha}{\omega_\nu}\right]\right\}e^{(-\alpha-j\omega_d)t} + \left(1 - j\frac{\alpha}{\omega_d}\right)\int_0^t \frac{d}{d\tau}\left\{\frac{Z[j\omega(\tau)]}{2}\right.$$

$$\left.\left[\left(\frac{\omega_d-\omega(\tau)}{\omega(\tau)}\right) + j\frac{\alpha}{\omega(\tau)}\right]\right\}e^{(-\alpha-j\omega_d)(t-\tau)}\, e^{j\int_0^\tau \omega(\xi)d\xi}\, d\tau$$

$$(12)$$

Equations (11) and (12) can be simplified to yield

$$K_1(t) = \left(1 + j\frac{\alpha}{\omega_d}\right)\frac{1}{2C}\,\frac{j(\omega_o-\omega_\nu)}{[\alpha+j(\omega_\nu-\omega_d)]\,[\alpha+j(\omega_o-\omega_d)]}\,e^{(-\alpha+j\omega_d)t}$$

$$+ \left(1 + j\frac{\alpha}{\omega_d}\right)\frac{1}{2C}\int_0^t \frac{j\omega'(\tau)}{[\alpha+j(\omega(\tau)-\omega_d)]^2}\,e^{(-\alpha+j\omega_d)(t-\tau)}\,e^{j\int_0^\tau \omega(\xi)d\xi}\,d\tau$$

$$(13)$$

$$K_2(t) = \left(1 - j\frac{\alpha}{\omega_d}\right)\frac{1}{2C}\,\frac{j(\omega_o-\omega_\nu)}{[\alpha+j(\omega_\nu+\omega_d)]\,[\alpha+j(\omega_o+\omega_d)]}\,e^{(-\alpha-j\omega_d)t}$$

$$+ \left(1 - j\frac{\alpha}{\omega_d}\right)\frac{1}{2C}\int_0^t \frac{j\omega'(\tau)}{[\alpha+j(\omega(\tau)+\omega_d)]^2}\,e^{(-\alpha-j\omega_d)(t-\tau)}\,e^{j\int_0^\tau \omega(\xi)d\xi}\,d\tau$$

where $\omega'(\tau) = \frac{d\omega(\tau)}{d\tau}$. Equation (10) clearly demonstrates that the response can be expressed as a quasi-stationary term plus a correction term.

That portion of the correction term involving ω_o and ω_ν

is due to the step change in frequency occurring at $t = 0$. When the slope of the instantaneous phase angle of the excitation is continuous at $t = 0$, $\omega_o = \omega_\nu$ and these terms go to zero. The remaining portions of the correction term reflect the departure of the system from quasi-stationary behavior. They arise because the system response is unable to build up or decay as fast as the quasi-stationary term would indicate. Note that only the first derivatives of $Z[j\omega(\tau)]$ and $\omega(\tau)$ appear in this part of the response. When the variations change slowly enough, the derivatives are small and the system behaves in a quasi-stationary manner.

A phasor interpretation of Equation (10) is often useful. Dividing through by $Z[j\omega(t)]\, e^{j\int_o^t \omega(\xi)d\xi}$, the normalized response is

$$e_o(t)_{\text{Normalized}} = 1 + \frac{K_1(t) + K_2(t)}{Z[j\omega(t)]}\, e^{-j\int_o^t \omega(\xi)d\xi} \tag{15}$$

The quasi-stationary term is now a constant unit phasor. As the correction phasor interferes with the constant phasor, the resultant phasor undergoes amplitude and frequency modulation. These modulations represent the departure from quasi-stationary response. (See Figure 3.)

As is often the case, once a solution has been obtained by a rather lengthy and involved procedure, the very same solution can be derived in a much shorter and less complicated manner. In the appendix the FM problem is solved for an n^{th} order system using only a single integration by parts of the appropriate convolution integral. The response again consists of a quasi-stationary term plus a correction term. Even for higher order systems only the first derivatives of the system function and hence of the instantaneous frequency of the excitation affect the correction term.

RESPONSE OF A TIME-VARYING PARALLEL RC CIRCUIT TO A FREQUENCY MODULATED SINUSOID

The circuit to be analyzed is given in Figure 4. The quasi-stationary impedance, $Z[j\omega_i,t]$, is defined to be the conventional steady-state impedance in which the parameters R, C, and ω are replaced by $R(t)$, $C(t)$, and $\omega_i(t)$ respectively. $Z[j\omega_i,t]$ should not be confused with Zadeh's[7] system function.

The latter is defined to be $H(j\omega,t) = y(t)/e^{j\omega t}$ where $y(t)$ is the system response to the input $e^{j\omega t}$. For $t < 0$, $R(t)$ and $C(t)$ are constant. At $t = 0$, the resistance and capacitance begin to vary in some arbitrary but specified manner. The circuit elements are always assumed to be positive. The excitation is identical to that of the previous problem and Figure 1 - b and Equations (1) and (2) still apply.

As before, the solution begins by making piece-wise constant approximations of the time-varying parameters. These are illustrated in Figure 5. During the time interval, $n\Delta\tau \leq t < (n + 1)\Delta\tau$, the parameters are given by $\hat{R}(t) = R_n$, $\hat{C}(t) = C_n$, $\hat{\omega}(t) = \omega_n$ and the circuit is characterized by a first-order differential equation with constant coefficients. In this interval the solution is given by

$$e_o(t) = \underbrace{Z[j\omega_n, n\Delta\tau]\, e^{j[\sum_{u=0}^{n-1} \omega_u\Delta\tau + \omega_n(t - n\Delta\tau)]}}_{\text{particular solution}} + \underbrace{K_n e^{-\alpha_n(t - n\Delta\tau)}}_{\text{Homogeneous solution}}$$

where

$$Z[j\omega_n, n\Delta\tau] = \frac{\frac{1}{C_n}}{\alpha_n + j\omega_n} \quad \text{and} \quad \alpha_n = \frac{1}{R_n C_n} \tag{16}$$

The constant K_n is evaluated by using the condition that the charge on the capacitor is continuous at $t = n\Delta\tau$. This results in the following recursion relation for K_n:

$$K_n = \begin{cases} Z[j\omega_\nu, 0] - Z[j\omega_o, 0] & n = 0 \\ \left\{\left[\frac{C_{n-1}}{C_n} Z[j\omega_{n-1}, (n-1)\Delta\tau] - Z[j\omega_n, n\Delta\tau]\right]\right\} e^{j[\sum_{u=0}^{n-1} \omega_u\Delta\tau]} + \frac{C_{n-1}}{C_n} K_{n-1} e^{-\alpha_{n-1}\Delta\tau} & n = 1, 2, 3, \ldots \end{cases} \tag{17}$$

Combining the recursion relation results in

$$K_n = \begin{cases} Z[j\omega_\nu, 0] - Z[j\omega_o, 0] & \\ \left\{\frac{C_o}{C_1} Z[j\omega_o, 0] - Z[j\omega_1, \Delta\tau]\right\} e^{j\omega_o \Delta\tau} + \frac{C_o}{C_1}\left\{Z[j\omega_\nu, 0] - Z[j\omega_o, 0]\right\} e^{-\alpha_o \Delta\tau} & \\ \left\{\frac{C_{n-1}}{C_n} Z[j\omega_{n-1}, (n-1)\Delta\tau] - Z[j\omega_n, n\Delta\tau]\right\} e^{j[\sum_{u=0}^{n-1} \omega_u \Delta\tau]} & n = o \\ + \left[\prod_{k=1}^{n} \frac{C_{k-1}}{C_\kappa}\right] \left\{Z[j\omega_\nu, 0] - Z[j\omega_o, 0]\right\} \left[\prod_{q=o}^{n-1} e^{-\alpha q\Delta\tau}\right] & n = 1 \\ + \sum_{\ell=1}^{n-1} \left[\prod_{k=\ell+1}^{n} \frac{C_{k-1}}{C_k}\right] \left\{\frac{C_{\ell-1}}{C_\ell} \quad Z[j\omega_{\ell-1}, (\ell-1)\Delta\tau] \right. & \\ & n=2,\ 3,\ 4,\ \ldots \\ \qquad \left. - Z[j\omega_\ell, \ell\Delta\tau]\right\} e^{j[\sum_{u=0}^{n} \omega_u \Delta\tau]} \left[\prod_{q=\ell}^{n-1} e^{-\alpha q\Delta\tau}\right] & \end{cases} \tag{18}$$

The solution for the continuous case is obtained by making a finer and finer approximation of the time-varying parameters in such a manner that $n\Delta\tau \to t$, $\ell\Delta\tau \to \tau$, $u\Delta\tau \to \xi$, $q\Delta\tau \to \sigma$, and $k\Delta\tau \to \delta$. The limiting operation yields

$$e_o(t) = \underbrace{Z[j\omega, t]\, e^{j\int_o^t \omega(\xi)d\xi}}_{\text{quasi-stationary term}} + \underbrace{K(t)}_{\substack{\text{correction}\\ \text{term}}} \tag{19}$$

where

$$K(t) = \frac{C_o}{C(t)}\left\{Z[j\omega_\nu, 0] - Z[j\omega_o, 0]\right\} e^{-\int_o^t \alpha(\sigma)\, d\sigma} \tag{20}$$

$$- \frac{1}{C(t)} \int_o^t \frac{d}{d\tau}\Big[C(\tau)\, Z[j\omega,\tau]\Big]\, e^{j\int_o^\tau \omega(\xi)d\xi}\, e^{-\int_\tau^t \alpha(\sigma)d\sigma}\, d\tau$$

The step change in frequency occurring at t = 0 is responsible for that portion of K(t) involving ω_o and ω_ν. The remainder of K(t) is the departure of the response from quasi-stationary behavior.

When the frequency of the excitation is constant for all time, Equations (19) and (20) reduce to

$$e_o(t) = Z[j\omega, t]\, e^{j\omega t} + K(t) \tag{21}$$

where

$$K(t) = -\frac{1}{C(t)} \int_0^t \frac{d}{d\tau}\Big[C(\tau)\, Z[j\omega,\tau]\Big]\, e^{j\omega\tau}\, e^{-\int_\tau^t \alpha(\sigma)d\sigma}\, d\tau \qquad (22)$$

Zadeh's system function, $H(j\omega,t)$, is obtained from Equation (21) by division with $e^{j\omega t}$. $H(j\omega,t)$ is of particular significance since it can be used to determine the response of the system to any arbitrary input having a Fourier transform by the relation

$$e_o(t) = \int_{-\infty}^{\infty} X'j\omega)\; H(j\omega,\, t)\; e^{j\omega t}\, \frac{d\omega}{2\pi} \qquad (23)$$

where $X(j\omega)$ is the Fourier transform of the input.

One may wonder whether it is possible to vary $R(t)$ and $C(t)$ simultaneously so that $K(t)$ is identically zero. This occurs when α is a constant (i.e. - when the resistance varies inversely proportional to the capacitance). The situation is not completely trivial since although the denominator of $Z[j\omega, t]$ is constant under this condition, the numerator varies inversely proportional to $C(t)$. Hence, it is possible to impedance level a parallel RC circuit as a function of time by appropriately varying R and C.

When R and C are constant for all time, Equations (19) and (20) reduce to

$$e_o(t) = Z[j\omega(t)]\, e^{j\int_0^t \omega(\xi)d\xi} + K(t) \qquad (24)$$

where

$$K(t) = \frac{1}{C}\, \frac{j(\omega_o - \omega_v)}{[\alpha + j\omega_v]\,[\alpha + j\omega_o]}\, e^{-\alpha t} + \frac{1}{C} \int_0^t \frac{j\omega'(\tau)}{[\alpha + j\omega(\tau)]^2}\, e^{-\alpha(t-\tau)}\, e^{j\int_0^t \omega(\xi)d\xi}\, d\tau \qquad (25)$$

Note the similarity between Equations (10, 13 - 14) and Equations (24 - 25). As pointed out in the appendix, this similarity carries through for circuits of higher order.

RESPONSE OF A TIME-VARYING SERIES RC CIRCUIT TO SINUSOIDAL EXCITATION OF CONSTANT FREQUENCY

Consider the circuit shown in Figure 6. The quasi-stationary admittance and the quasi-stationary voltage transfer ratio are defined from the corresponding steady-state quantities in the usual manner. The excitation is a constant frequency sinusoid which is applied to the network at t = -∞. The resistance and capacitance variations of Figure 4-b are also assumed for this problem.

Application of the method of solution used in the previous problems gives the following expressions for $i(t)$ and $e_o(t)$:

$$i(t) = \underbrace{Y[j\omega,t]\, e^{j\omega t}}_{\text{quasi-stationary term}} + \alpha(t) \underbrace{\int_0^t \frac{d}{d\tau}\Big[C(\tau)\, G[j\omega,\tau]\Big]\, e^{j\omega\tau}\, e^{-\int_\tau^t \alpha(\sigma)d\sigma}\, d\tau}_{\text{correction term}} \qquad (26)$$

$$e_o(t) = \underbrace{G[j\omega,t]\, e^{j\omega t}}_{\text{quasi-stationary term}} - \frac{1}{C(t)} \underbrace{\int_0^t \frac{d}{d\tau}\Big[C(\tau)\, G[j\omega,\tau]\Big]\, e^{j\omega\tau}\, e^{-\int_\tau^t \alpha(\sigma)d\sigma}\, d\tau}_{\text{correction term}} \qquad (27)$$

Once again, the responses are characterized by an interference phenomenon between the appropriate quasi-stationary term and a correction term. The correction terms, including that of Equation (22), have the property that when the time-varying elements change slowly enough, they become negligible due to the derivatives involved. Because of the factor $e^{j\omega\tau}$ in the integrands, they also become negligible when ω is exceedingly large. This is reasonable since, in such a case, many cycles of excitation will occur before the time-varying elements have changed appreciably. Thus, the network will appear as a constant paramenter network to the excitation and the quasi-stationary term will predominate.

In deriving Equations (26) and (27) the excitation was applied at t = -∞ so that Zadeh's system function could be obtained simply by dividing with $e^{j\omega t}$. A problem of more practical interest is one in which the excitation is suddenly applied at t = 0. For this situation the voltage response can be shown to be

$$e_o(t) = G[j\omega, t]\, e^{j\omega t} - \frac{1}{C(t)} \int_0^t \frac{d}{d\tau}\Big[C(\tau)\, G[j\omega,\tau]\Big]\, e^{j\omega\tau}\, e^{-\int_\tau^t \alpha(\sigma)d\sigma}\, d\tau$$

$$+ \Big[E_o - G[j\omega, 0]\Big] \frac{C_o}{C(t)}\, e^{-\int_\tau^t \alpha(\sigma)d\sigma} \qquad (28)$$

where E_o is the initial voltage across the capacitor. Note that Equation (28) is identical to Equation (27) except for an added term which arises in order to satisfy the initial conditions at t = 0. This term is the basis function of the system (i.e. - the solution of the homogeneous differential equation). It is anticipated that the basis functions for higher order systems can be derived in a similar fashion.

A specific example is now considered to demonstrate the utility of phasors in predicting the behavior of time-varying circuits. The example was first suggested by Cruz[8]. The excitation is given by

$$e_s(t) = u_{-1}(t)e^{j\omega t} \tag{29}$$

where $u_{-1}(t)$ is a unit step occuring at t = 0. The capacitance is constant and is designated by C. The initial voltage, E_o, is zero. R varies with time according to the relation

$$R(t) = \frac{t + 1}{C} \tag{30}$$

Hence,

$$\alpha(t) = \frac{1}{R(t)C} = \frac{1}{t + 1} \tag{31}$$

Equation (28) is applicable and yields, after some manipulation,

$$e_o(t) = \frac{1}{j\omega(t + 1)} \quad [e^{j\omega t} - 1] \tag{32}$$

The response can be interpreted as the interaction between a fixed phasor and a rotating phasor as shown in Figure 7. At t = 0, the two phasors cancel and $e_o(t) = 0$, the initial voltage across the capacitor. As the moving phasor rotates at the frequency ω, the resultant phasor wobbles back and forth while growing and decreasing in magnitude. The wobbling of the resultant readily explains why the output of a time-varying circuit contains many frequency components not included in the input. In fact, in this particular example, the phasor model clearly indicates that the instantaneous frequency of the response will undergo severe frequency spikes each time the moving phasor rotates through 2π radians and the envelope of the response goes to zero. The envelope of the response goes to zero whenever $t = 2\pi k/\omega$,

$k = 1, 2, 3, \ldots$. At a specific instant of time, t_o, Equation (32) can be regarded as a function of frequency. The phasor model then shows that the envelope of the response will be zero at all frequencies for which $\omega = 2\pi\kappa/t_o$, $k = 1, 2, 3, \ldots$.

CONCLUDING REMARKS

This paper has attempted to bring the realm of time-invariant circuit theory into the domain of time-varying circuits. It was shown that the interference between a quasi-stationary term and a correction term is basic to the understanding of both time-invariant filters excited by frequency-modulated wave forms and time-varying circuits. It is suggested that time-varying circuits be viewed in terms of this interference phenomenon. It has also been shown that time-invariant circuit concepts can be introduced into the analysis of time-varying circuits.

The response of first order RL circuits can be obtained from the RC circuits discussed in this paper using the principle of duality. The response of higher order time-varying circuits is presently under investigation. Preliminary results indicate that the solutions will be of the same form as those presented in this paper. The problem is complicated by the fact that higher order circuits can become unstable[9,10]. It is not yet clear whether the method of solution proposed will yield valid results for such cases.

APPENDIX

RESPONSE OF AN Nth ORDER TIME-INVARIANT SYSTEM TO A FREQUENCY-MODULATED SIGNAL

Consider a system function

$$H(s) = \frac{A_m s^m + \ldots + A_o}{(s+s_1)(s+s_2)\ \ldots\ (s+s_n)} \quad , m < n,\ s_i \neq s_j \qquad \text{(A-1)}$$

which has the partial fraction expansion

$$H(s) = \frac{K_1}{s+s_1} + \frac{K_2}{s+s_2} + \ldots + \frac{K_n}{s+s_n} = \sum_{y=1}^{n} \frac{K_y}{s+s_y} \qquad \text{(A-2)}$$

The impulse response is given by

$$h(t) = K_1 e^{-s_1 t} + K_2 e^{-s_2 t} + \ldots + K_n e^{-s_n t} = \sum_{y=1}^{n} K_y e^{-s_y t} \qquad \text{(A-3)}$$

Denote the input by x(t) and the output by w(t). Let x(t) be given by

$$x(t) = \begin{cases} e^{j\omega_\nu t} & t \leq 0 \\ e^{j\theta(t)} = e^{j\int_0^t \omega(\xi)d\xi} & t > 0 \end{cases}$$

The response of the linear system is

$$w(t) = \int_{-\infty}^{t} h(t - \tau)\ x(\tau)\ d\tau$$

$$= \int_{-\infty}^{0} \sum_{y=1}^{n} K_y e^{-s_y(t-\tau)} e^{j\omega_\nu \tau} d\tau + \int_0^t \sum_{y=1}^{n} K_y e^{-s_y(t-\tau)} e^{j\theta(\tau)} d\tau \tag{A-4}$$

Assuming that $s_y = \alpha_y + j\omega_y$ where $\alpha_y > 0$,

$$w(t) = \sum_{y=1}^{n} \frac{K_y}{s_y + j\omega_\nu} e^{-s_y t} + \sum_{y=1}^{n} \int_0^t K_y e^{-s_y(t-\tau)} e^{j\theta(\tau)} d\tau \tag{A-5}$$

The formula for integration by parts is

$$\int_a^b u dv = uv \Big|_a^b - \int_a^b v du \tag{A-6}$$

Applying to the y^{th} integral in Equation (A-5), let

$$u_y = \frac{1}{s_y + j\omega(\tau)}$$

$$\frac{du_y}{d\tau} = \frac{d}{d\tau} \frac{1}{s_y + j\omega(\tau)} = -\frac{j\omega'(\tau)}{[s_y + j\omega(\tau)]^2}$$

$$v_y = K_y e^{-s_y(t-\tau)} e^{j\theta(\tau)} \tag{A-8}$$

$$\frac{dv_y}{d\tau} = K_y [s_y + j\omega(\tau)] e^{-s_y(t-\tau)} e^{j\theta(\tau)}$$

It follows that

$$\int_0^t K_y e^{-s_y(t-\tau)} e^{j\theta(\tau)} d\tau = \frac{K_y}{[s_y + j\omega(t)]} e^{j\theta(t)}$$

$$- \frac{K_y}{s_y + j\omega_0} e^{-s_y t} - \int_0^t \frac{d}{d\tau}\left[\frac{K_y}{s_y + j\omega(\tau)}\right] e^{-s_y(t-\tau)} e^{j\theta(\tau)} d\tau \tag{A-8}$$

Substitution of Equation (A-8) into Equation (A-5) results in

$$\begin{aligned} w(t) = {} & \sum_{y=1}^{n} \frac{K_y}{s_y + j\omega(t)} e^{j\theta(t)} \\ & + \sum_{y=1}^{n} \frac{K_y}{s_y + j\omega_\nu} e^{-s_y t} - \sum_{y=1}^{n} \frac{K_y}{s_y + j\omega_o} e^{-s_y t} \\ & - \sum_{y=1}^{n} \int_o^t \frac{d}{d\tau}\left[\frac{K_y}{s_y + j\omega(\tau)}\right] e^{-s_y(t-\tau)} e^{j\theta(\tau)} d\tau \end{aligned} \tag{A-9}$$

This reduces to

$$\begin{aligned} w(t) = {} & Z[j\omega(t)] e^{j\theta(t)} \\ & + \sum_{y=1}^{n} K_y \frac{j(\omega_o - \omega_\nu)}{(s_y + j\omega_\nu)(s_y + j\omega_o)} e^{-s_y t} \\ & + \sum_{y=1}^{n} \int_o^t \frac{j\omega'(\tau)}{[s_y + j\omega(\tau)]^2} e^{-s_y(t-\tau)} e^{j\theta(\tau)} d\tau \end{aligned} \tag{A-10}$$

Equation (A-10) specializes to Equations (10, 13-14) of the text for the case of a single-tuned circuit.

REFERENCES

1. J. R. Carson and T. C. Fry, "Variable-Frequency Electric Circuit Theory," Bell Sys. Tech. J., vol. 16, pp. 513-540, October, 1937.

2. B. van der Pol, "The Fundamental Principles of Frequency Modulation," J. IEEE (London), vol. 93, pt. 3, pp. 153-158, May, 1946.

3. F.L.H.M. Stumpers, "Distortion of Frequency-Modulated Signals in Electrical Networks," Commun. News, vol. 9, pp. 82-92, April, 1948.

4. E. J. Baghdady, "Theory of low-Distortion Reproduction of FM Signals in Linear Systems," Trans, IRE, PGCS, vol. CT-5, pp. 202-214, September, 1958.

5. D. T. Hess, "Transmission of FM Signals Through Linear Filters," Proc. of NEC, vol. 18, October, 1962.

6. J. J. Hupert, "Transient REsponse of Narrow-Band Networks to Angle-Modulated Signals," Proc. of NEC, vol. 18, Oct., 1962.

7. L. A. Zadeh, "Frequency Analysis of Variable Networks," Proc. of IRE, vol. 38, No. 3, pp. 291-299, March, 1950.

8. J. B. Cruz, Jr., "Linear Systems with Time-Varying Components." Proc. of NEC, vol. 17, pp. 71-80, October, 1961.

9. C. A. Desoer and A. Paige, "Linear Time-Varying Networks: Stable and Unstable," IEEE International Convention, March, 1963.

10. S. Darlington, "Linear Time-Varying Circuits-Matrix Manipulations, Power relations, and Some Bounds on Stability," to be published in Bell Sys. Tech. J.

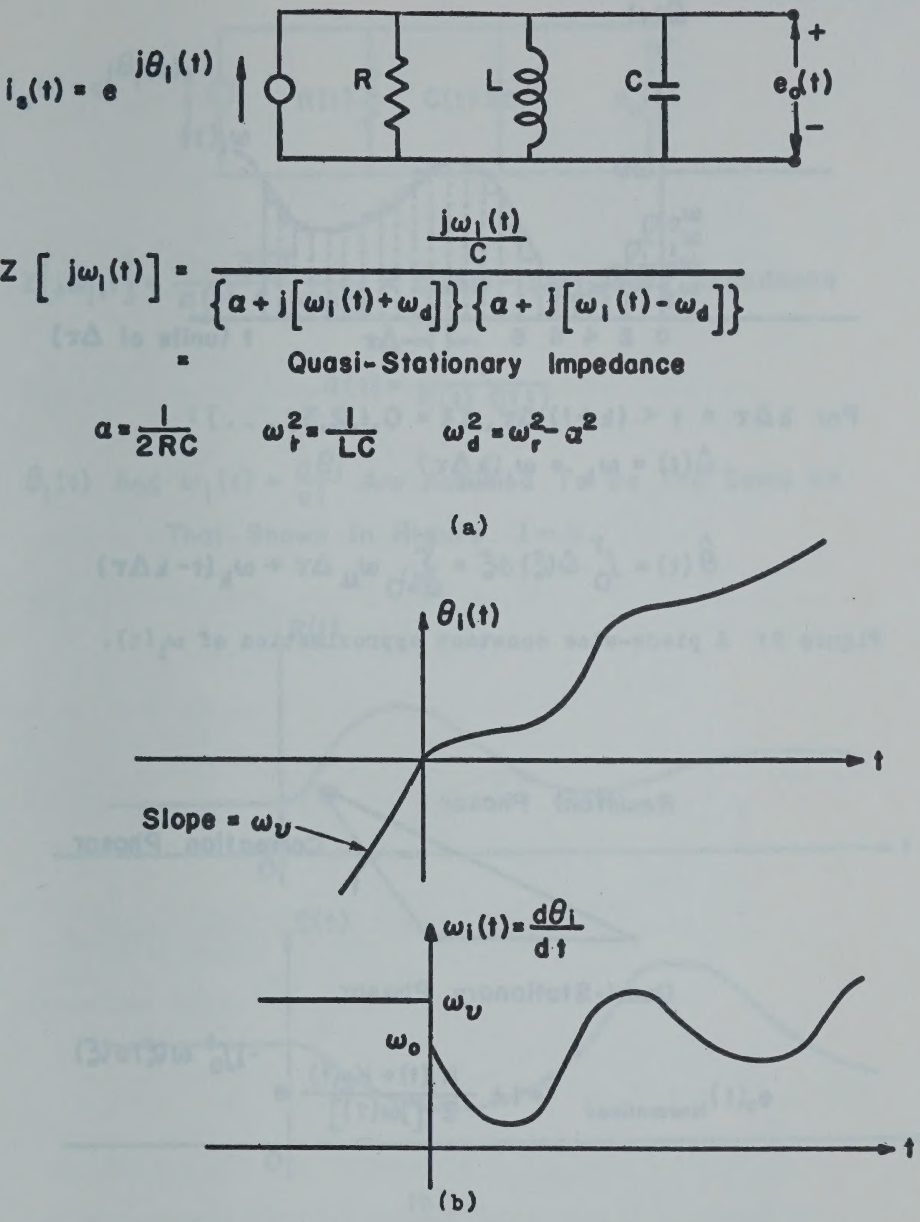
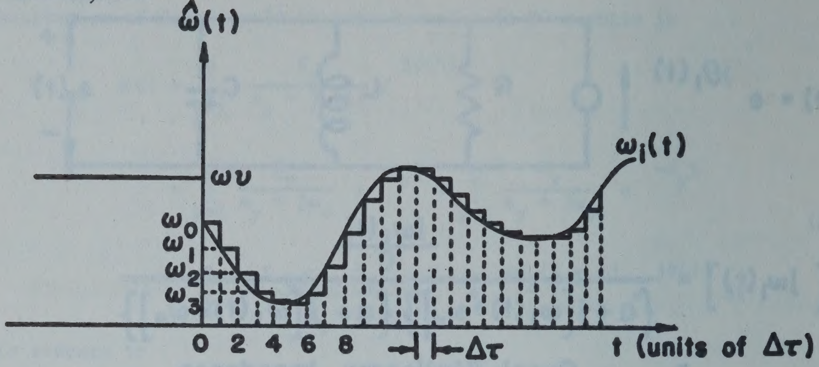

$i_s(t) = e^{j\theta_1(t)}$
R
L
C
+
$e_o(t)$
−
$Z[j\omega_1(t)] = \dfrac{\frac{j\omega_1(t)}{C}}{\{\alpha + j[\omega_1(t) + \omega_d]\}\{\alpha + j[\omega_1(t) - \omega_d]\}}$
= Quasi-Stationary Impedance
$\alpha = \frac{1}{2RC}$ $\omega_r^2 = \frac{1}{LC}$ $\omega_d^2 = \omega_r^2 - \alpha^2$
(a)
$\theta_1(t)$
t
Slope = ω_v
$\omega_1(t) = \frac{d\theta_1}{dt}$
ω_v
ω_o
t
(b)

Figure 1: (a) Single-tuned circuit with definition of quasi-stationary impedance (b) Instantaneous phase angle and instantaneous frequency of the excitation.

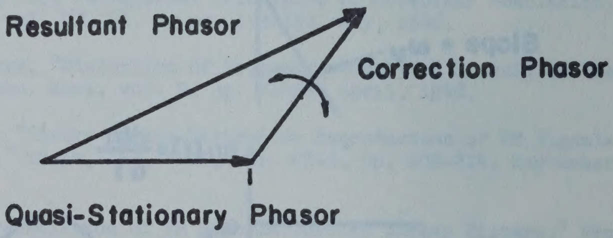


For $k\Delta\tau \leq t < (k+1)\Delta\tau$, $(k = 0,1,2,3,\ldots)$:

$$\hat{\omega}(t) = \omega_k = \omega_i(k\Delta\tau)$$

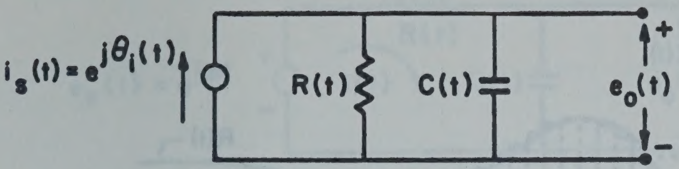
$$\hat{\theta}(t) = \int_0^t \hat{\omega}(\xi)\,d\xi = \sum_{\mu=0}^{k-1} \omega_\mu \Delta\tau + \omega_k (t - k\Delta\tau)$$

Figure 2: A piece-wise constant approximation of $\omega_i(t)$.


Resultant Phasor
Correction Phasor
Quasi-Stationary Phasor

$$e_o(t)_{\text{Normalized}} = 1 + \frac{K_1(t) + K_2(t)}{Z\left[j\omega(t)\right]}\, e^{-j\int_0^t \omega(\xi)\,d(\xi)}$$

Figure 3: Phasor representation of Equation (15). The amplitude and frequency modulation of the resultant phasor represents the departure from quasi-stationary response.

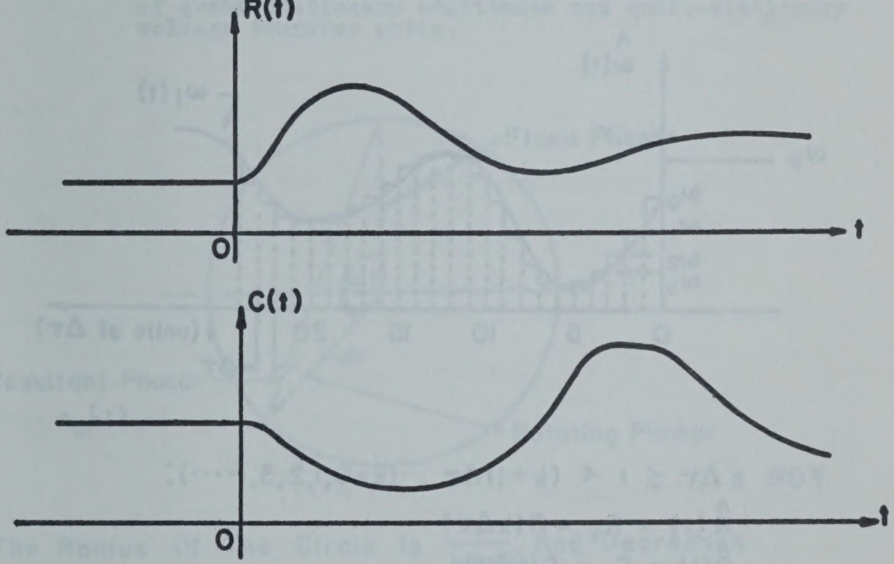


$$Z\left[j\omega_i, t\right] = \frac{\frac{1}{C(t)}}{\alpha(t) + j\omega_i(t)}$$ = Quasi - Stationary Impedance

$$\alpha(t) = \frac{1}{R(t)\ C(t)}$$

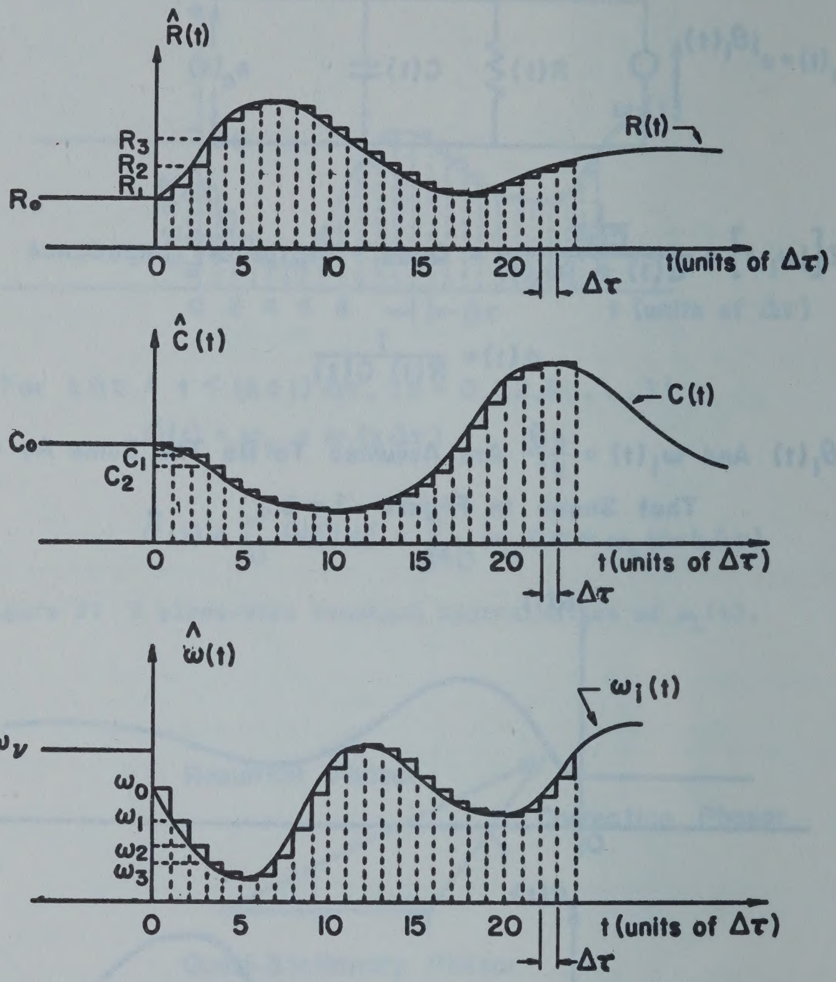
$\theta_i(t)$ And $\omega_i(t) = \frac{d\theta_i}{dt}$ Are Assumed To Be The Same As That Shown In Figure 1 - b.

(a)



(b)

Figure 4: (a) Time-varying parallel RC circuit with definition of quasi-stationary impedance, $\theta_i(t)$, and $\omega_i(t)$. (b) Plots of R(t) and C(t) versus time.



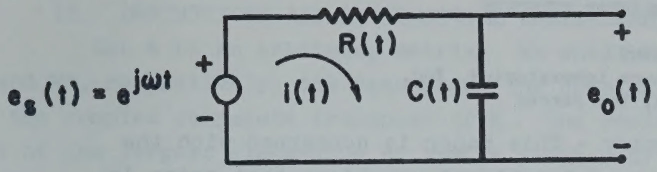
FOR $k\,\Delta\tau \le t < (k+1)\Delta\tau$, $(k=0,1,2,3,\cdots)$:

$$\hat{R}(t) = R_k = R(k\Delta\tau)$$

$$\hat{C}(t) = C_k = C(k\Delta\tau)$$

$$\hat{\omega}(t) = \omega_k = \omega_1(k\Delta\tau)$$

Figure 5: Piece-wise constant approximations of R(t), C(t), and $\omega_1(t)$.

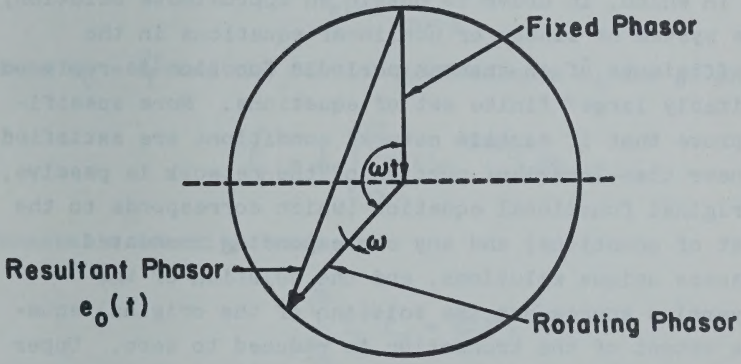

$e_s(t) = e^{j\omega t}$
+
-
R(t)
i(t)
C(t)
+
$e_0(t)$
-

$$Y\left[j\omega, t\right] = \frac{j\omega \;\frac{1}{R(t)}}{\alpha(t) + j\omega} = \text{Quasi-Stationary Admittance}$$

$$G\left[j\omega, t\right] = \frac{\alpha(t)}{\alpha(t) + j\omega} = \text{Quasi-Stationary Voltage Transfer Ratio}$$

$$\alpha(t) = \frac{1}{R(t)\,C(t)}$$

Figure 6: Time-varying series RC Circuit with definitions of quasi-stationary admittance and quasi-stationary voltage transfer ratio.


Fixed Phasor
ωt
ω
Resultant Phasor
$e_0(t)$
Rotating Phasor

The Radius Of The Circle Is $\frac{1}{\omega(t+1)}$ And Decreases In Magnitude With Increasing Time And/Or Increasing Frequency.

Figure 7: A phasor interpretation of Equation (32).

ON TRUNCATION TECHNIQUES IN THE APPROXIMATE ANALYSIS OF PERIODICALLY TIME-VARYING NONLINEAR NETWORKS

I. W. SANDBERG
Bell Telephone Laboratories, Inc.
Murray Hill, New Jersey

Summary - This paper is concerned with the properties of functional equations that arise in the analysis of periodically time-varying nonlinear networks. In particular, we establish in a completely rigorous manner the validity of two well-known truncation techniques for obtaining approximate solutions. Upper bounds on the truncation error are presented.

I. INTRODUCTION

This paper is concerned with the properties of functional equations that arise in the analysis of electrical networks which contain linear time-invariant elements and a finite, but arbitrary, number of resistors that are periodically time-varying or nonlinear. In particular, we establish the mathematical legitimacy of two well-known truncation techniques in which, in order to obtain an approximate solution, an infinite system of linear or nonlinear equations in the Fourier coefficients of an unknown periodic function is replaced with a "suitably large" finite set of equations. More specifically, we prove that if certain natural conditions are satisfied and the linear time-invariant portion of the network is passive, both the original functional equation (which corresponds to the infinite set of equations) and any corresponding truncated version possess unique solutions, and the solution of the truncated version approaches the solution of the original equation as the extent of the truncation is reduced to zero. Upper bounds on the truncation error are presented.

Some definitions and mathematical preliminaries are considered in Section II. In Section III we consider the analysis of electrical networks containing time-varying nonlinear resistors and describe the usual truncation techniques for obtaining approximate solutions. Section IV is concerned with mathematical results relating to functional equations of the type introduced earlier.

II. DEFINITIONS AND MATHEMATICAL PRELIMINARIES

Let M be an arbitrary matrix. We shall denote by M', $\overline{M}$, and M^*, respectively, the transpose, the complex conjugate, and the complex conjugate transpose of M. The positive square-root of the largest eigenvalue of M^*M is denoted by $\Lambda\{M\}$.

The space of complex N-vector valued functions of t is denoted by $\mathcal{H}_N$. Let T be a real positive constant and let

$$\mathcal{K}_N = \left\{ f \,|\, f \in \mathcal{H}_N,\ f(t) = f(t+T),\ \int_0^T f^*f \, dt < \infty \right\}.$$

The norm of $f = (f_1, f_2, \ldots, f_N)' \in \mathcal{K}_N$ is denoted by $\|f\|$ and is defined by

$$\|f\|^2 = \frac{1}{T} \int_0^T f^*f \, dt.$$

With this norm $\mathcal{K}_N$ is a Banach space.[1]

If $f \in \mathcal{K}_N$,

$$f = \underset{K\to\infty}{\text{l.i.m.}} \sum_{n=-K}^{K} \hat{f}_n e^{in\omega_0 t}$$

where $\omega_0 = \frac{2\pi}{T}$ and the Fourier coefficients $\hat{f}_n$ are given by

$$\hat{f}_n = \frac{1}{T} \int_0^T f(t) e^{-in\omega_0 t} dt.$$

Parseval's identity reads:

$$\sum_{-\infty}^{\infty} \hat{f}_n{}^* \hat{f}_n = \|f\|^2.$$

Two elements of $\mathcal{K}_N$, f and h, are equivalent if $\|f-h\| = 0$.

Let Σ denote a set of integers. The projection operator $\underline{P}$ is the linear mapping of $\mathcal{K}_N$ into itself defined by

$$\begin{aligned} \frac{1}{T} \int_0^T (\underline{P}f) e^{-ip\omega_0 t} dt &= \hat{f}_p, \quad p \in \Sigma \\ &= 0, \quad p \notin \Sigma \end{aligned}$$

where $f \in \mathcal{K}_N$.

In accordance with the usual notation the norm

$$\sup_{\substack{f \in \mathcal{K}_N \\ \|f\| = 1}} \|\underline{Q}f\|$$

of a linear operator $\underline{Q}$ defined on $\mathcal{K}_N$ is denoted by $\|\underline{Q}\|$.

The symbol $\mathcal{L}_{1R}$ denotes the space of real-valued absolutely integrable functions defined on $(-\infty,\infty)$. We take as the definition of the Fourier transform of $x(t) \in \mathcal{L}_{1R}$:

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt.$$

The symbol $\underline{I}$ is used throughout to denote the identity operator on $\mathcal{K}_N$, and 1_N denotes the identity matrix of order N.

The set of real-valued functions is denoted by $\mathcal{R}$.

Let $\mathcal{E} = [\Theta,\rho]$ be an arbitrary metric space.[1] A mapping $\underline{A}$ of the space $\mathcal{E}$ into itself is said to be a contraction if there exists a number $k < 1$ such that

$$\rho(\underline{A}x,\underline{A}y) \leq k\rho(x,y)$$

for any two elements $x,y \in \Theta$. The contraction-mapping fixed-point theorem[1] is basic to much of the subsequent discussion. It states that every contraction-mapping defined in a complete metric space $\mathcal{E}$ has one and only one fixed point (i.e., there exists a unique element $z \in \Theta$ such that $\underline{A}z = z$). Furthermore $z = \lim_{n\to\infty} \underline{A}^n z_0$, where z_0 is an arbitrary element of Θ.

III. ANALYSIS OF NETWORKS CONTAINING TIME-VARYING NONLINEAR RESISTORS; TRUNCATION TECHNIQUES

Consider the analysis of an electrical network containing independent voltage and current sources, linear (not necessarily lumped) time-invariant elements, and N time-varying nonlinear resistors. As is well known, all currents and voltages within the network can be determined by routine techniques once the voltages (or currents) associated with the time-varying nonlinear elements are known.

Let $i_n(t)$ and $v_n(t)(1 \leq n \leq N)$ denote, respectively, the current through and voltage across the nth time-varying nonlinear resistor and let this resistor introduce the constraint:

$$i_n(t) = g_n(v_n(t),t)$$

in which $g_n(\cdot,\cdot)$ is real-valued function. Let $i = (i_1,i_2,\ldots,i_N)'$, $v = (v_1,v_2,\ldots,v_N)'$ and define $\underline{G}$ by

$$\underline{G}v = (g_1(v_1,t),g_2(v_2,t),\ldots,g_N(v_N,t))'$$

so that $i = \underline{G}v$.

We assume that the Thévenin-type relation between v and i introduced by the linear time-invariant portion of the network is

$$e(t) = v(t) + \underline{Z}i(t) \tag{1}$$

in which e(t) is independent of v(t) and i(t), and $\underline{Z}$ is the linear operator defined by

$$\underline{Z}i(t) = \int_{-\infty}^{\infty} z(t-\tau)i(\tau)d\tau,$$

where z(t) is an N×N (not necessarily symmetric) matrix of functions of t.

From (1) and $i = \underline{G}v$,

$$v + \underline{Z}\underline{G}v = e. \tag{2}$$

Assumption I: It is assumed throughout that every element of z(t) belongs to $\mathfrak{L}_{1R}$.

Let

$$Z(\omega) = \int_{-\infty}^{\infty} z(t)e^{-i\omega t}dt.$$

Assumption I and the Riemann-Lebesgue lemma imply that $Z(\omega) \to 0$ as $|\omega| \to \infty$, by which we mean that every element of $Z(\omega)$ approaches zero as $|\omega| \to \infty$. Thus in this sense the linear time-invariant portion of the network is assumed to behave at infinity as a network with each port shunted by a capacitor.

If $\underline{Z}$ corresponds to a passive network, $[Z(\omega) + Z^*(\omega)]$ is nonnegative definite for all real ω.

The mathematical arguments presented in Section IV do not require that $z(t) = 0$ for $t < 0$. That is, it is not necessary to assume that $\underline{Z}$ is a causal operator.

We next discuss two problems that are of considerable engineering interest.

3.1 The Linear Problem

Suppose that all N resistors are linear, positive, and periodically time-varying with the same period T; that is, suppose that for all $v_n(t)$

$$g_n(v_n(t),t) = g_n(1,t)v_n(t) \qquad n = 1,2,\dots,N$$

in which $g_n(1,t) > 0$ and $g_n(1,t) = g_n(1,t+T)$.

The basic problem here is to determine the N-vector of voltages v(t) when $e(t) = e_0 e^{i\omega_1 t}$, where e_0 is an arbitrary constant N-vector and ω_1 is an arbitrary real constant. It is well known that the linearity and periodically time-varying character of the system imply that if a unique solution v(t) exists, $v(t) = u(t)e^{i\omega_1 t}$ in which $u(t) = u(t+T)$. Equation (2) implies that u(t) must satisfy:

$$e_0 = u + \underline{Z}_1\underline{G}u, \tag{3}$$

in which $\underline{Z}_1$ is obtained from $\underline{Z}$ by replacing z(t) by $z(t)e^{-i\omega_1 t}$.

A well-known technique for obtaining an approximate solution of (3) is to solve a "suitably large" finite set of linear equations obtained by truncating the infinite system of equations involving the Fourier coefficients of u(t) implied by (3). That is, it is assumed that (3) possesses one and only one solution belonging to $\mathcal{K}_N$ and that it is approximately $\tilde{u}$, the exact solution of

$$e_0 = \tilde{u} + \underline{PZ_1G}\tilde{u} \tag{4}$$

for some finite Σ. The set Σ, which is ordinarily chosen in accordance with some intuitive criterion, must be such that $Z(\omega_1 - p\omega_0)$ is "suitably small" if $p \notin \Sigma$. Although this technique is of considerable practical utility, it appears that the following basic questions have not been discussed in the literature:

(i) Does $\mathcal{K}_N$ contain one and only one solution of (3)?

(ii) Does the truncated set of equations possess a unique solution? More specifically, for a given $\underline{P}$ (i.e., for a given Σ), does $\mathcal{K}_N$ contain one and only one solution of (4)?

(iii) Assuming that (3) and (4) possess unique solutions belonging to $\mathcal{K}_N$, is the solution of (4) close to that of (3)? In particular, what is an upper bound on the error[†] involved and for an arbitrary positive constant γ, does there exist a finite subset Ξ of the integers such that the mean-square truncation error $\|u-\tilde{u}\|^2$ is at most γ whenever $\Xi \subseteq \Sigma$?

3.2 The Nonlinear Problem

Let us now consider a different but analogous situation. Suppose that all N resistors are time-invariant but nonlinear. That is, suppose that for all t

$$g_n(v_n(t),t) = g_n(v_n(t),0) \qquad n = 1,2,\ldots,N.$$

Assume also that there exist two positive constants σ_1 and σ_2 such that

$$\sigma_1(x_1-x_2) \leq g_n(x_1,0) - g_n(x_2,0) \leq \sigma_2(x_1-x_2)$$

$(n = 1,2,\ldots,N)$ for all real $x_1 \geq x_2$.

Here an important problem is the determination of the N-vector v(t) when e(t) is a given real-valued periodic function with period T. The problem is often treated by assuming that there exists a unique exact solution $v \in \mathcal{K}_N \cap \mathcal{R}$, and, as in the linear problem, choosing a finite Σ such that there is reason to believe that $\tilde{v}$, the exact solution of

$$e = \tilde{v} + \underline{PZG}\tilde{v}, \tag{5}$$

is an acceptable approximation to v. We assume here that $\underline{P}$ is chosen so that it maps $\mathcal{K}_N \cap \mathcal{R}$ into itself. That is, we assume here that $n \in \Sigma$ if $-n \in \Sigma$.

With regard to $\tilde{v}$, the Fourier coefficients that correspond to frequencies $n\omega_0$ for $n \notin \Sigma$ are equal to the corresponding coefficients of e(t), while the remaining coefficients satisfy a finite set of nonlinear equations.[††]

It is clear that the questions analogous to those raised in connection with the linear problem are equally pertinent here. Indeed, the present situation perhaps appears to be even more uncertain because of the nonlinear character of the problem.

[†] Questions of this type have been considered for a special class of infinite systems of equations, the so-called "regular" systems.[2]

[††] This is essentially the describing function technique.

3.3 Preview of the Results

Section IV is concerned with mathematical properties of functional equations that are somewhat more general than those considered in the last two subsections. Assuming merely that $[Z(\omega) + Z^*(\omega)]$ is nonnegative definite for all real ω (which is the case when the linear time-invariant portion of the network is passive), the results in Section IV answer all of the questions raised. In particular, it is proved that for either the linear or nonlinear problem, $\mathcal{K}_N$ ($\mathcal{K}_N \cap \mathcal{R}$ for the nonlinear problem) contains one and only one solution, any corresponding finite set of truncated equations possesses a unique solution, and for any $\gamma > 0$ there exists a finite subset Ξ of the integers such that the mean-square truncation error is at most γ whenever $\Xi \subseteq \Sigma$. Explicit bounds on the truncation error are given.

IV. MATHEMATICAL RESULTS

Definition I: Let $\{Q_n\}_{-\infty}^{\infty}$ denote a countable set of N×N complex constant matrices such that

$$\sup_n \Lambda\{Q_n\} < \infty. \tag{6}$$

Let $\underset{\sim}{Q}$ be a bounded linear mapping of $\mathcal{K}_N$ into itself with the property that if $f \in \mathcal{K}_N$ and $h = \underset{\sim}{Q}f$, $\hat{h}_n = Q_n\hat{f}_n$, in which $\hat{f}_n$ and $\hat{h}_n$, respectively, are the nth Fourier coefficients of f and h. (Since

$$\sum_{-\infty}^{\infty} \hat{h}_n^*\hat{h}_n = \sum_{-\infty}^{\infty} \hat{f}_n^* Q_n^* Q_n \hat{f}_n \leq (\sup_n \Lambda\{Q_n\})^2 \sum_{-\infty}^{\infty} \hat{f}_n^*\hat{f}_n < \infty,$$

the Riesz-Fischer theorem implies that $\underset{\sim}{Q}$ is completely defined on $\mathcal{K}_N$ by $\{Q_n\}_{-\infty}^{\infty}$.)

In view of Assumption I, the class of operators satisfying Definition I includes the case in which[†]

$$\underset{\sim}{Q}f = \int_{-\infty}^{\infty} z(t-\tau)e^{-i\omega_1(t-\tau)} f(\tau)d\tau, \quad f \in \mathcal{K}_N$$

for any real ω_1. Here $Q_n = Z(n\omega_0+\omega_1)$.

The following preliminary result is proved in the Appendix.

[†] It is sufficient to prove this statement for the case in which N = 1. This is done in Ref. 3.

Theorem I:

(i) $\|\underline{Q}\| = \sup_n \Lambda\{Q_n\}$.

(ii) If $[1_N+Q_n]^{-1}$ exists for all n and $\sup_n \Lambda\{[1_N+Q_n]^{-1}\} < \infty$, the operator $(\underline{I}+\underline{Q})$ possesses a bounded inverse on $\mathcal{K}_N$ and

$$\|(\underline{I}+\underline{Q})^{-1}\underline{Q}\| = \sup_n \Lambda\{[1_N+Q_n]^{-1}Q_n\}.$$

Definition II: Let α and β be real constants such that $\beta \geq \alpha$, and, when $[1_N + \frac{1}{2}(\alpha+\beta)Q_n]^{-1}$ exists for all n, let

$$r = \tfrac{1}{2}(\beta-\alpha)\sup_n \Lambda\{[1_N + \tfrac{1}{2}(\alpha+\beta)Q_n]^{-1}Q_n\}.$$

The following theorem, which relates to the applicability of Theorems III and IV, is needed to establish the fact that the results in this section answer the questions raised in Section III, assuming merely that the linear time-invariant portion of the network is passive.

Theorem II: Suppose that $\alpha > 0$ and $[Q_n+Q_n^*]$ is nonnegative definite for all n. Then $[1_N + \frac{1}{2}(\alpha+\beta)Q_n]^{-1}$ exists for all n, $\sup_n \Lambda\{[1_N + \frac{1}{2}(\alpha+\beta)Q_n]^{-1}\} \leq 1$, and $r < 1$.

Proof:

Let $R_n = \frac{1}{2}(\alpha+\beta)Q_n$. Since

$$[1_N+R_n][1_N+R_n^*] = 1_N + R_n + R_n^* + R_nR_n^*$$

is positive definite, $\det\{[1_N+R_n][1_N+R_n^*]\} \neq 0$. Hence $\det[1_N+R_n] \neq 0$ for all n.

The quantity $(\Lambda\{[1_N + \frac{1}{2}(\alpha+\beta)Q_n]^{-1}\})^2$ is the largest eigenvalue of $(1_N+R_n+R_n^*+R_nR_n^*)^{-1}$, or, what is the same thing, the reciprocal of the smallest eigenvalue of $(1_N+R_n+R_n^*+R_nR_n^*)$. Because $(R_n+R_n^*+R_nR_n^*)$ is nonnegative definite, the smallest eigenvalue of $(1_N+R_n+R_n^*+R_nR_n^*)$ is at least unity. Thus

$$\sup_n \Lambda\{[1_N + \tfrac{1}{2}(\alpha+\beta)Q_n]^{-1}\} \leq 1.$$

Consider now the statement regarding r. In terms of α, β, and R_n,

$$r = (\beta-\alpha)(\beta+\alpha)^{-1}\sup_n \Lambda\{[1_N+R_n]^{-1}R_n\}.$$

It suffices to show that

$$\Lambda\{[1_N+R_n]^{-1}R_n\} \leq 1$$

for all n. For an arbitrary N×N matrix M,

$$(\Lambda\{M\})^2 = \sup \frac{x^*M^*Mx}{x^*x}$$

as x varies over all nonzero constant N-vectors. Using $[1_N+R_n]^{-1}R_n = 1_N - [1_N+R_n]^{-1}$,

$$R_n^*[1_N+R_n]^{-1*}[1_N+R_n]^{-1}R_n$$

$$= 1_N - [1_N+R_n]^{-1*} - [1_N+R_n]^{-1} + [1_N+R_n]^{-1*}[1_N+R_n]^{-1}$$

$$= 1_N - [1_N+R_n]^{-1*}[1_N+R_n+R_n^*][1_N+R_n]^{-1}. \qquad (7)$$

Since $[1_N+R_n]^{-1*}[1_N+R_n+R_n^*][1_N+R_n]^{-1}$ is nonnegative definite, it is evident from (7) that

$$\frac{x^*R_n^*[1_N+R_n]^{-1*}[1_N+R_n]^{-1}R_nx}{x^*x} \leq 1$$

for any nonzero constant N-vector x and any n. This proves the theorem.

The following result and its analog, Theorem IV, relating to the nonlinear case and the space $\mathcal{K}_N \cap \mathcal{R}$, are the key results of the paper.

Theorem III: Let $a_1(t), a_2(t), \ldots, a_N(t)$ denote real-valued measurable periodic functions of t, with period T, such that there exist real constants α and β with the property that

$$\alpha \leq a_n(t) \leq \beta \qquad n = 1,2,\ldots,N$$

for almost all t in [0,T]. Let $\underline{A}$ be the operator defined by

$$\underline{A}h = (a_1(t)h_1(t), a_2(t)h_2(t), \ldots, a_N(t)h_N(t))', \qquad h \in \mathcal{K}_N$$

and let $\underline{Q}$ be as defined earlier.

Suppose that

(i) $[1_N + \frac{1}{2}(\alpha+\beta)Q_n]^{-1}$ exists for all n

(ii) $\sup_n \Lambda\{[1_N + \frac{1}{2}(\alpha+\beta)Q_n]^{-1}\} < \infty$

(iii) $r < 1$.

Then for any $g \in \mathcal{K}_N$ there exists a unique $f \in \mathcal{K}_N$ such that $g = f + \underline{QA}f$. In fact, $f = \lim_{n\to\infty} f^{(n)}$, where

$$f^{(n+1)} = - [\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}[\underline{A} - \tfrac{1}{2}(\alpha+\beta)\underline{I}]f^{(n)} + [\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}g$$

and $f^{(0)}$ is an arbitrary element of $\mathcal{K}_N$. The nth approximation satisfies:

$$\|f^{(n)}-f\| \leq \frac{r^n}{1-r}\,\|f^{(1)}-f^{(0)}\|.$$

Proof:

According to Theorem I [with $\underline{Q}$ replaced with $\frac{1}{2}(\alpha+\beta)\underline{Q}$], assumptions (i) and (ii) imply that $[\underline{I} + \frac{1}{2}(\alpha+\beta)\underline{Q}]$ possesses a bounded inverse on $\mathcal{K}_N$. Hence the functional equation $g = f + \underline{Q}\underline{A}f$ can be written as $f = \underline{S}f$, where

$$\underline{S}f = - [\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}[\underline{A} - \tfrac{1}{2}(\alpha+\beta)\underline{I}]f + [\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}g.$$

It is evident that $\underline{S}$ is a mapping of $\mathcal{K}_N$ into itself. Let us consider a condition under which $\underline{S}$ is in fact a contraction-mapping of $\mathcal{K}_N$ into itself.

Denote by $\underline{\tilde{A}}$ the operator $\underline{A} - \frac{1}{2}(\alpha+\beta)\underline{I}$, let $f_a, f_b \in \mathcal{K}_N$ and observe that

$$\begin{aligned}\|\underline{S}f_a-\underline{S}f_b\| &= \|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}\{\underline{\tilde{A}}f_a-\underline{\tilde{A}}f_b\}\|\\ &\leq \|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}\|\cdot\|\underline{\tilde{A}}f_a-\underline{\tilde{A}}f_b\|.\end{aligned}$$

Since

$$|a_n(t) - \tfrac{1}{2}(\alpha+\beta)| \leq \tfrac{1}{2}(\beta-\alpha) \qquad n = 1,2,\ldots,N$$

it follows that

$$\|\underline{\tilde{A}}f_a-\underline{\tilde{A}}f_b\| \leq \tfrac{1}{2}(\beta-\alpha)\|f_a-f_b\|. \tag{8}$$

Thus $\underline{S}$ is a contraction if

$$\tfrac{1}{2}(\beta-\alpha)\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}\| < 1. \tag{9}$$

According to Theorem I, the left-hand side of (9) is r, and, by assumption, $r < 1$. In view of the contraction-mapping fixed-point theorem, this establishes the existence and uniqueness of the solution f(t) and the fact that it can be determined in accordance with the stated iteration procedure.

The upper bound on $\|f^{(n)}-f\|$ follows directly from the fact that f can be written as

$$f = f^{(0)} + \sum_{j=0}^{\infty} [f^{(j+1)} - f^{(j)}]$$

in which for all $j \geq 1$

$$\|f^{(j+1)} - f^{(j)}\| = \|\underline{S}f^{(j)} - \underline{S}f^{(j-1)}\| \leq r\|f^{(j)} - f^{(j-1)}\|.$$

Corollary III(a): Suppose that the hypotheses of Theorem III are satisfied. Then

$$\|f\| \leq \frac{1}{1-r} \|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1} g\|.$$

Proof:

Set $n = 0$ and $f^{(0)} = 0$ in the upper bound for $\|f - f^{(n)}\|$.

Corollary III(b): Suppose that the hypotheses of Theorem III are satisfied. Then for any $g \in \mathcal{K}_N$ there exists a unique $\tilde{f} \in \mathcal{K}_N$ such that $g = \tilde{f} + \underline{PQA}\tilde{f}$.

Proof:

The hypotheses of Theorem III remain satisfied if $\underline{Q}$ is replaced with $\underline{PQ}$.

Corollary III(c): Suppose that the hypotheses of Theorem III are satisfied and that $\tilde{f} \in \mathcal{K}_N$ satisfies $g = \tilde{f} + \underline{PQA}\tilde{f}$. Then

$$\begin{aligned} \|f - \tilde{f}\| &\leq \frac{1}{1-r} \|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}(\underline{I}-\underline{P})[\underline{A}\tilde{f} - \tfrac{1}{2}(\alpha+\beta)(\tilde{f}-g)]\| \\ &= \frac{1}{1-r} \|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}(\underline{I}-\underline{P})\underline{A}\tilde{f}\|. \end{aligned}$$

Proof:

We note first that in view of Theorem I, assumptions (i) and (ii) in Theorem III, and the definition of $\underline{P}$, $[I + \tfrac{1}{2}(\alpha+\beta)\underline{PQ}]$ possesses a bounded inverse on $\mathcal{K}_N$.

With $n = 0$ and $f^{(0)} = \tilde{f}$, the upper bound for $\|f^{(n)} - f\|$ yields:

$$\begin{aligned} \|f - \tilde{f}\| \leq \frac{1}{1-r} \|&[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}[\underline{A}\tilde{f} - \tfrac{1}{2}(\alpha+\beta)\tilde{f}] - [\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}g \\ &- [\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{PQ}]^{-1}\underline{PQ}[\underline{A}\tilde{f} - \tfrac{1}{2}(\alpha+\beta)\tilde{f}] + [\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{PQ}]^{-1}g\|. \end{aligned} \tag{10}$$

The corollary follows directly from (10) and the identities:

$$\underline{PQ} = \underline{QP}, \quad (\underline{I}-\underline{P})\tilde{f} = (\underline{I}-\underline{P})g,$$

$$[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{PQ}]^{-1} = [\underline{I}-\underline{P}] + [\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{P}.$$

Corollary III(d): Suppose that the hypotheses of Theorem III are satisfied and let $\tilde{f}$ satisfy $g = \tilde{f} + \underline{PQA}\tilde{f}$. Let

$$q = \tfrac{1}{2}(\beta-\alpha) \sup_{n \,\varepsilon\, \Sigma} \Lambda\{[1_N + \tfrac{1}{2}(\alpha+\beta)Q_n]^{-1}Q_n\}$$

$$s = \sup_{n \not\varepsilon\, \Sigma} \Lambda\{[1_N + \tfrac{1}{2}(\alpha+\beta)Q_n]^{-1}Q_n\}.$$

Then

$$\|f-\tilde{f}\| \leq \frac{s}{1-r}\{\tfrac{1}{2}(\beta-\alpha)(1-q)^{-1}\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{P}g\|$$

$$+ \tfrac{1}{2}[|\alpha+\beta| + (\beta-\alpha)(1-q)^{-1}]\|(\underline{I}-\underline{P})g\|\}.$$

Proof:

From Corollary III(c),

$$\|f-\tilde{f}\| \leq \frac{1}{1-r}\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}(\underline{I}-\underline{P})\|\cdot\|(\underline{I}-\underline{P})[\underline{A}\tilde{f} - \tfrac{1}{2}(\alpha+\beta)(\tilde{f}-g)]\|. \tag{11}$$

Clearly

$$\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}(\underline{I}-\underline{P})\| = \|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}(\underline{I}-\underline{P})]^{-1}\underline{Q}(\underline{I}-\underline{P})\|$$

$$= \sup_{n \not\varepsilon\, \Sigma} \Lambda\{[1_N + \tfrac{1}{2}(\alpha+\beta)Q_n]^{-1}Q_n\} = s \tag{12}$$

and

$$\|(\underline{I}-\underline{P})[\underline{A}\tilde{f} - \tfrac{1}{2}(\alpha+\beta)(\tilde{f}-g)]\| \leq \|\underline{A}\tilde{f} - \tfrac{1}{2}(\alpha+\beta)\tilde{f}\| + \tfrac{1}{2}|\alpha+\beta|\cdot\|(\underline{I}-\underline{P})g\|$$

$$\leq \tfrac{1}{2}(\beta-\alpha)\|\tilde{f}\| + \tfrac{1}{2}|\alpha+\beta|\cdot\|(\underline{I}-\underline{P})g\|.$$

From Corollary III(a), with $\underline{Q}$ replaced with $\underline{PQ}$,

$$\|\tilde{f}\| \leq \frac{1}{1-q}\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{PQ}]^{-1}g\| \leq \frac{1}{1-q}\{\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{P}g\| + \|(\underline{I}-\underline{P})g\|\}.$$

Thus

$$\|(\underline{I}-\underline{P})[\underline{A}\tilde{f} - \tfrac{1}{2}(\alpha+\beta)(\tilde{f}-g)]\| \leq \tfrac{1}{2}(\beta-\alpha)(1-q)^{-1}\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{P}g\|$$

$$+ \tfrac{1}{2}[|\alpha+\beta| + (\beta-\alpha)(1-q)^{-1}]\|(\underline{I}-\underline{P})g\|. \tag{13}$$

The result stated in the corollary is a direct consequence of (11), (12), and (13).

Remarks:

Suppose that $Q_n \rightarrow 0$ as $n \rightarrow \infty$, as is the case in the network problems in Section III. Then, in view of the definition of s and the fact that

$$\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{P}g\| \le \|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}g\|$$

$$\|(\underline{I}-\underline{P})g\| \le \|g\| \qquad \frac{1}{1-q} \le \frac{1}{1-r},$$

it follows that for any $\gamma > 0$ there exists a finite subset Ξ of the integers such that $\|f-\tilde{f}\|^2 \le \gamma$ whenever $\Xi \subseteq \Sigma$.

In many cases of interest $(\underline{I}-\underline{P})g = 0$, with the result that the upper bound on the truncation error reduces to simply:

$$\|f-\tilde{f}\| \le \frac{1}{2}\,\frac{(\beta-\alpha)s}{(1-r)(1-q)}\,\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{P}g\|.$$

4.1 Results for the Analogous Nonlinear Case and the Space $\mathcal{H}_N \cap \mathcal{R}$.

The result analogous to Theorem III for the nonlinear case and the space $\mathcal{H}_N \cap \mathcal{R}$ is

Theorem IV: Let $b_1(x,t), b_2(x,t), \ldots, b_N(x,t)$ denote real-valued functions of the real variables x and t, which for fixed x are continuous and periodic in t with period T. Assume that there exist real constants α and β such that for $n = 1,2,\ldots,N$

$$\alpha \le \frac{b_n(x_1,t) - b_n(x_2,t)}{x_1-x_2} \le \beta$$

for all real $x_1 \ne x_2$ and all t in $[0,T]$. Let $\underline{B}$ denote the operator defined by

$$\underline{B}h = (b_1(f_1,t), b_2(f_2,t), \ldots, b_N(f_N,t))', \qquad h \in \mathcal{H}_N \cap \mathcal{R}$$

and let $\underline{Q}$ be as defined earlier.

Suppose that

(i) $Q_n = \overline{Q}_n$ for all n
(ii) $[1_N + \frac{1}{2}(\alpha+\beta)Q_n]^{-1}$ exists for all n
(iii) $\sup_n \Lambda\{[1_N + \frac{1}{2}(\alpha+\beta)Q_n]^{-1}\} < \infty$
(iv) $r < 1$.

Then for any $g \in \mathcal{H}_N \cap \mathcal{R}$ there exists a unique $f \in \mathcal{H}_N \cap \mathcal{R}$ such that $g = f + \underline{QB}f$. In fact, $f = \lim_{n\to\infty} f^{(n)}$ where

$$f^{(n+1)} = -[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}[\underline{B} - \tfrac{1}{2}(\alpha+\beta)\underline{I}]f^{(n)} + [\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}g$$

and $f^{(0)}$ is an arbitrary element of $\mathcal{H}_N \cap \mathcal{R}$. The nth approximation satisfies

$$\|f^{(n)}-f\| \le \frac{r^n}{1-r}\,\|f^{(1)}-f^{(0)}\|.$$

Proof:

With the usual norm, $\mathcal{H}_N \cap \mathcal{R}$ is a Banach space. Thus, with essentially a single modification, the argument used to establish Theorem III suffices. Specifically, let $\underline{\tilde{B}} = \underline{B} - \frac{1}{2}(\alpha+\beta)\underline{I}$. Then since[†]

$$|b_n(x_1,t) - b_n(x_2,t) - \tfrac{1}{2}(\alpha+\beta)(x_1-x_2)| \leq \tfrac{1}{2}(\beta-\alpha)|x_1-x_2|,$$

$$\|\underline{\tilde{B}}f_a-\underline{\tilde{B}}f_b\| \leq \tfrac{1}{2}(\beta-\alpha)\|f_a-f_b\|$$

for any $f_a,f_b \in \mathcal{H}_N \cap \mathcal{R}$, which is the analog of (8).

The results analogous to Corollaries III(a) through III(d) are stated below. The proofs are omitted since arguments essentially the same as those used earlier suffice.

Corollary IV(a): Suppose that the hypotheses of Theorem IV are satisfied. Let u be the image of the zero element of $\mathcal{H}_N \cap \mathcal{R}$ under the mapping $\underline{QB}$. Then

$$\|f\| \leq \frac{1}{1-r}\,\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}(g-u)\|.$$

Corollary IV(b): Suppose that the hypotheses of Theorem IV are satisfied and that $n \in \Sigma$ if $-n \in \Sigma$. Then for any $g \in \mathcal{H}_N \cap \mathcal{R}$ there exists a unique $\tilde{f} \in \mathcal{H}_N \cap \mathcal{R}$ such that $g = \tilde{f} + \underline{PQB}\tilde{f}$.

Corollary IV(c): Suppose that the hypotheses of Theorem IV are satisfied, that $n \in \Sigma$ if $-n \in \Sigma$, and that $\tilde{f} \in \mathcal{H}_N \cap \mathcal{R}$ satisfies $g = \tilde{f} + \underline{PQB}\tilde{f}$. Then

$$\|f-\tilde{f}\| \leq \frac{1}{1-r}\,\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}(\underline{I}-\underline{P})[\underline{B}\tilde{f} - \tfrac{1}{2}(\alpha+\beta)(\tilde{f}-g)]\|$$

$$= \frac{1}{1-r}\,\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{Q}(\underline{I}-\underline{P})\underline{B}\tilde{f}\|.$$

Corollary IV(d): Suppose that the hypotheses of Theorem IV are satisfied, that $n \in \Sigma$ if $-n \in \Sigma$, and that $\tilde{f}$ satisfies $g = \tilde{f} + \underline{PQB}\tilde{f}$. Let w be the image of the zero element of $\mathcal{H}_N \cap \mathcal{R}$ under $\underline{B}$ and let q and s be the constants defined in Corollary III(d). Then

$$\|f-\tilde{f}\| \leq \frac{s}{1-r}\,\{\tfrac{1}{2}(\beta-\alpha)(1-q)^{-1}\|[\underline{I} + \tfrac{1}{2}(\alpha+\beta)\underline{Q}]^{-1}\underline{P}(g-\underline{Q}w)\|$$

$$+ \tfrac{1}{2}[|\alpha+\beta| + (\beta-\alpha)(1-q)^{-1}]\|(\underline{I}-\underline{P})g\| + \|(\underline{I}-\underline{P})w\|\}.$$

[†] The assumed t-continuity of $b_n(x,t)$ implies that $\underline{B}0 \in \mathcal{H}_N \cap \mathcal{R}$, and is used otherwise only to establish that $b_n[y(t),t]$ is a measurable function of t whenever y is measurable. Much weaker assumptions suffice for these purposes.

V. FINAL REMARK

In this paper the basic space of N-vector valued functions is $\mathcal{K}_N$. Arguments not essentially different from those already presented lead to entirely analogous results for the space

$$\left\{ f \mid f \,\varepsilon\, \mathcal{H}_N, \int_{-\infty}^{\infty} f^* f \, dt < \infty \right\}.$$

APPENDIX

Proof of Theorem I

The norm of $\underline{Q}$ is $\sup\{\|\underline{Q}f\|;\ f \,\varepsilon\, \mathcal{K}_N,\ \|f\| = 1\}$. Using Parseval's identity,

$$\|\underline{Q}f\|^2 = \sum_{-\infty}^{\infty} \hat{f}_n^{\,*} Q_n^{\,*} Q_n \hat{f}_n .$$

Thus, $\|\underline{Q}\| \leq \sup\limits_n \Lambda\{Q_n\}$, and since it is clear that there exists an $f \,\varepsilon\, \mathcal{K}_N$ such that $\|f\| = 1$ and $\|\underline{Q}f\| > \sup\limits_n \Lambda\{Q_n\} - \delta$, where δ is an arbitrary positive number, $\|\underline{Q}\| = \sup\limits_n \Lambda\{Q_n\}$.

Consider now the invertibility of the operator $(\underline{I}+\underline{Q})$ when $\sup\limits_n \Lambda\{[1_N+Q_n]^{-1}\} < \infty$. Let $w \,\varepsilon\, \mathcal{K}_N$ with Fourier coefficients $\hat{w}_n$. The hypothesis implies that

$$\sum_{-\infty}^{\infty} \hat{w}_n^{\,*} [1_N+Q_n]^{-1*} [1_N+Q_n]^{-1} \hat{w}_n < \infty .$$

Thus, according to the Riesz-Fischer theorem, there exists a $z \,\varepsilon\, \mathcal{K}_N$ with Fourier coefficients

$$\hat{z}_n = [1_N+Q_n]^{-1} \hat{w}_n .$$

Parseval's identity implies that z satisfies the equation $(\underline{I}+\underline{Q})z = w$ (in the sense that $\|(\underline{I}+\underline{Q})z - w\| = 0$) and that

$$\|z\| \leq \sup_n \Lambda\{[1_N+Q_n]^{-1}\} \|w\| .$$

Thus, $(\underline{I}+\underline{Q})$ possesses a bounded inverse on $\mathcal{K}_N$.

The expression for $\|(\underline{I}+\underline{Q})^{-1}\underline{Q}\|$ given in Theorem I follows from arguments similar to those used to obtain the expression for $\|\underline{Q}\|$.

REFERENCES

1. Kolmogorov, A. N. and Fomin, S. V., Elements of the Theory of Functions and Functional Analysis, New York, Graylock Press, 1957.

2. Kantorovich, L. V. and Krylov, V. I., Approximate Methods of Higher Analysis, New York, Interscience Publishers, 1958.

3. Sandberg, I. W., On the Response of Nonlinear Control Systems to Periodic Input Signals, to be published.

NATURAL FREQUENCIES OF TUNNEL DIODES*

V. W. CHANG and I. T. FRISCH
Department of Electrical Engineering and
Electronics Research Laboratory
University of California
Berkeley 4, California

I. Introduction

In 1960 Kinariwala[1] found the limitations on the natural frequencies of a tunnel diode imbedded in a passive network, with the tunnel diode treated as a negative resistor in parallel with a capacitor. This work was extended by Golosman and Newcomb[2] by treating the same problem with a resistor or an inductor (but not both) in series with Kinariwala's equivalent circuit for the tunnel diode. In this paper all possible natural frequencies in the right half of the complex frequency plane are found for a tunnel diode imbedded in a passive network, using the complete linear equivalent circuit for the tunnel diode. This equivalent circuit consists of a negative resistor in parallel with a capacitor and the combination in series with an inductor and a resistor.

The approach is to specialize Desoer's and Kuh's[3] conditions for an n-port network to be "active at p_0" to one-port networks and then apply these conditions to the equivalent circuit of the tunnel diode. Desoer and Kuh's conditions are first briefly stated and then simplified.

II. Networks Active at p_0

A real, linear time-invariant n-port network N is said to be active at the complex frequency p_0 with $\operatorname{Re} p_0 \geq 0$, i.e., the real part of p not less than zero, if there exists a passive n-port N_p such that the combined network $N{-}N_p$, formed by imbedding N in N_p as in Fig. 1, can support a mode of the form $\underline{i}(t) = \underline{i}e^{p_0 t}$ where $\underline{i} = \operatorname{col}(\underline{i}_1, \underline{i}_2, \cdots \underline{i}_n)$ is some fixed column vector. In other words, N is active at p_0 if there is an N_p such that the network $N{-}N_p$ has a short circuit natural frequency at p_0. The point p_0 is called an active point and the region in the complex frequency plane containing all points at which N is active is called the "active region." Desoer and Kuh[3] have shown that for a lumped network possessing an impedance matrix which is analytic at p_0, Condition I below is necessary for N to be active at p_0. Newcomb[4]

---

* This work was supported by the Air Force Office of Scientific Research of the Office of Aerospace Research; the Department of the Army, Army Research Office; and the Department of the Navy, Office of Naval Research, under Grant No. AF-AFOSR 139-63.

has shown that Condition I is necessary and sufficient when N is a one- or two-port network.

Condition I $q_+ \leq 0$

where

$$q_+ = \min_{||I||=1} I^{*t} \underline{Z}_H(p_0)\underline{I} \qquad \text{if } \omega_0 = 0 \text{ or } \sigma_0 = 0 \qquad (1a)$$

$$q_+ = \min_{||I||=1} I^{*t} \underline{Z}_H(p_0)\underline{I} + \frac{\sigma_o}{|p_0|} |\underline{I}^t \underline{Z}(p_0)\underline{I}| \qquad \text{if } \omega_o \neq 0 \text{ and } \sigma_o > 0 \qquad (1b)$$

and

$p_0 = \sigma_0 + j\omega_0$

$\underline{Z}(p)$ is the impedance matrix of N

$\underline{Z}_H(p)$ is the Hermitian part of $\underline{Z}(p)$

I is a fixed column vector corresponding to the port currents

* denotes the complex conjugate of a matrix

t denotes the transpose of a matrix

$||I||$ denotes $I^{*t} I$.

If N is not analytic at p_0, then N has a natural frequency at p_0 and N is active at p_0 for an open circuit termination.

Since we will be dealing with a tunnel diode treated as a one-port circuit, we must consider Condition I restricted to the one-port case. For a one-port network with input impedance $Z(p)$, $Z_H(p) = \mathrm{Re}Z(p)$ and $|\underline{Z}(p)| = (\mathrm{Re}\ Z)^2 + (\mathrm{Im}\ Z)^2$, where Re Z and Im Z denote the real and imaginary parts of Z, respectively. Condition I therefore reduces to

Condition II $q_+ \leq 0$

where $q_+ = \mathrm{Re}\ Z(p_0)$ if $\omega = 0$ or $\sigma = 0$ (2a)

$$\text{and } q_+ = \mathrm{Re}\ Z(p_0) + |\mathrm{Re}\ Z(p_0)| \sqrt{\frac{1 + (\mathrm{Im}\ Z(p_0)/\mathrm{Re}\ Z(p_0))^2}{1 + (\omega_0/\sigma_0)^2}} \qquad (2b)$$

if $\omega > 0$ and $\sigma > 0$.

The region $\omega < 0$ need not be considered, since we assume $Z^*(p) = Z(p^*)$. Thus the active region is symmetric about the real frequency axis.

If $\omega = 0$ or $\sigma = 0$, Condition II becomes

$$\operatorname{Re} Z(p_0) \leq 0 \tag{3}$$

If $\omega > 0$, $\sigma > 0$ Condition II becomes

$$\operatorname{Re} Z(p_0) \leq 0 \tag{4a}$$

and

$$(\operatorname{Im} Z(p_0)/\operatorname{Re} Z(p_0))^2 \leq (\omega_0/\sigma_0)^2 \tag{4b}$$

or since $\operatorname{Re} Z(p_0) \leq 0$ and $\omega > 0$ is assumed, we finally have

$$\operatorname{Re} Z(p_0) \leq 0 \tag{5a}$$

$$\sigma \operatorname{Im} Z(p_0) - \omega_0 \operatorname{Re} Z(p_0) \geq 0 \tag{5b}$$

$$\sigma \operatorname{Im} Z(p_0) + \omega_0 \operatorname{Re} Z(p_0) \leq 0 \tag{5c}$$

Equations (3) and (5a, b, c) are the conditions to be applied to the tunnel diode equivalent circuit. The degenerate case of $\operatorname{Re} Z = 0$ is included in (5), since $q_+ \leq 0$, requires $\operatorname{Im} Z = 0$ and thus $q_+ = 0$ for this case.

III. Tunnel-Diode Equivalent Circuit

A reasonable small-signal, equivalent circuit for a tunnel diode is given in Fig. 2a, where L_s represents the lead inductance, R_s the bulk resistance, C the diode barrier capacitance, and -g the negative tunneling conductance. After using the impedance normalization factor g and the frequency normalization factor g/C, the normalized equivalent circuit to be used in the remainder of the paper is given in Fig. 2b.

The impedance Z(p) for the normalized equivalent circuit is

$$Z(p) = r + \ell p + \frac{1}{p-1} = r + \ell\sigma + \frac{\sigma - 1}{(\sigma-1)^2+\omega^2} + j\left[\ell\omega - \frac{\omega}{(\sigma-1)^2+\omega^2}\right] \tag{6}$$

or

$$\operatorname{Re} Z(p) = r + \ell\sigma + \frac{\sigma - 1}{(\sigma-1)^2+\omega^2} \tag{7a}$$

$$\operatorname{Im} Z(p) = \ell\omega - \frac{\omega}{(\sigma-1)^2+\omega^2} \tag{7b}$$

IV. Tunnel Diode Active Region

We can now proceed to determine the active region for a single tunnel diode. Substituting (7a) and (7b) into (5a, b, c) we have, respectively, (8)(9) and (10) as follows. (Note that (8) also corresponds to (7a) and (7b) substituted into (3).)

$$\ell\sigma^3+\ell\sigma\omega^2+(r-2\ell)\sigma^2+r\omega^2+(1-\ell-2r)\sigma+r-1 \le 0 \tag{8}$$

$$r\sigma^2 + r\omega^2 + 2(1-r)\sigma + r-1 \le 0 \tag{9}$$

$$2\ell\sigma^3 + 2\ell\sigma\omega^2 + (r-4\ell)\sigma^2+r\omega^2+(2\ell-2r)\sigma+r-1 \le 0 \tag{10}$$

Rearranging, we have:

$$\text{From (8) } \omega^2 \le \frac{1-\sigma}{r+\ell\sigma} - (\sigma-1)^2 \tag{11}$$

$$\text{From (9) } \omega^2 \le 1/r\,(1-2\sigma)-(\sigma-1)^2 \tag{12}$$

$$\text{From (10) } \omega^2 \le \frac{1}{2\ell\sigma+r} - (\sigma-1)^2 \tag{13}$$

The regions of the complex frequency plane satisfying (11), (12) and (13) then consist of the points in the first quadrant of the complex frequency plane enclosed by the axes and the values of ω satisfying (14) (15) and (16), respectively.

$$\omega^2 = \frac{1-\sigma}{r+\ell\sigma} - (\sigma-1)^2 \tag{14}$$

$$\omega^2 = 1/r\,(1-2\sigma) - (\sigma-1)^2 \tag{15}$$

$$\omega^2 = \frac{1}{2\ell\sigma+r} - (\sigma-1)^2 \tag{16}$$

Values of ω^2 satisfying (14), (15) and (16) will be denoted by ω_I^2, ω_{II}^2, and ω_{III}^2, respectively. Then a point σ_o on the real axis will be an active point if and only if $\omega_I^2 \ge 0$ for that σ_o. A point ω_o on the imaginary axis will be an active point if and only if ω_o^2 is not greater than the value of ω_I^2 at $\sigma = 0$. A point $p_0 = \sigma_0 + j\omega_0$ in the first quadrant (not including the axes) is an active point provided that ω_I^2, ω_{II}^2 and ω_{III}^2 are all positive at σ_o, and provided that $\omega_0^2 \le \text{Min}\,(\omega_I^2, \omega_{II}^2, \omega_{III}^2)$ at $\sigma = \sigma_0$. For a given value of σ, ω_I^2 or ω_{II}^2 or ω_{III}^2 is said to be "dominant" or to be the "dominant restriction" if it is equal to Min $(\omega_I^2, \omega_{II}^2, \omega_{III}^2)$. Equation (14), (15) or (16) is said to be the "dominant curve" or the "dominant equation" if it yields the dominant restriction. It is then easily shown that the relative values of ω_I^2, ω_{II}^2, and ω_{III}^2 and, hence, the dominant restrictions for different ranges of r, ℓ and σ are as given in Table I.

The problem of finding the active regions in the first quadrant for the tunnel diode has therefore been reduced to examining the dominant restrictions for different ranges of r, ℓ, and σ. The active regions for the fourth quadrant are then found by reflecting the region in the first quadrant about the real axis. In Section V, the region is found over which $\omega_I^2 \geq 0$. Since ω_I^2 determines the active region on the real and imaginary axis, these regions are then found and given in Table 2. In Section VI the region of non-negative ω_{II}^2 is determined. From Table 1 we see that for $r \geq \ell$, ω_{II}^2 is dominant. We can therefore determine the active region (not including the axes) from ω_{II}^2. Combining this region with the active region on the axes found in Section V, the complete active region for $r \geq \ell$ is then given in Table 3. In Section VII, the region of non-negative ω_{III}^2 is determined. Combining the results of Sections V, VI and VII, the active region for $r < \ell$ is then given in Table 5.

V. Region of Non-Negative ω_I^2 and Active Region for $\omega = 0$ or $\sigma = 0$.

We consider (14), which yields the region of activity on the real and imaginary axes. We first identify the regions over which $\omega_I^2 \geq 0$. If $\sigma > 1$, then $\omega_I^2 < 0$. Hence we can limit our investigation to the range $o < \sigma \leq 1$. If $\sigma = 1$, then $\omega_I^2 \geq 0$ corresponds to a single point, (1, 0) on the σ, axis. This point, in fact, corresponds to the open-circuit natural frequency of the tunnel diode. Next, consider $\sigma < 1$. Rearranging (14) we have

$$\omega_I^2 = \frac{1-\sigma}{r+\ell\sigma}\left[1-(1-\sigma)(r+\ell\sigma)\right] \tag{17}$$

$\frac{1-\sigma}{r+\ell\sigma}$ is non-negative, for $0 \leq \sigma \leq 1$, since r and ℓ are assumed positive. In order that ω_I^2 be positive, we therefore must have

$$F(\sigma) \equiv 1-(1-\sigma)(r+\ell\sigma) \geq 0 \tag{18}$$

or

$$F(\sigma) = \left[\sigma - 1/2(1-r/\ell)\right]^2 + \left[1/\ell - 1/4(1+r/\ell)^2\right] \geq 0 \tag{19}$$

We then have three cases to consider, as follows.

Case I. If $1/\ell - 1/4\,(1+r/\ell)^2 > 0$, then $F(\sigma)$ is always non-negative for all $\sigma < 1$. This condition is equivalent to $r < 2\sqrt{\ell} - \ell$.

Case II. If $1/\ell - 1/4(1+r/\ell)^2 = 0$, then

(a) For $r \geq \ell$, $\omega_I^2 \geq 0$ for $0 \leq \sigma \leq 1$

(b) For $r < \ell$, $\omega_I^2 \geq 0$ for $1/2(1-r/\ell) < \sigma \leq 1$

Case III. If $1/\ell - 1/4\,(1 + r/\ell)^2 < 0$, rearrange $F(\sigma)$ in factored form

$$F(\sigma) = [\,\sigma - 1/2(1 - r/\ell) + 1/2\sqrt{(1 + r/\ell)^2 - 4/\ell}\,]\,[\,\sigma - 1/2(1 - r/\ell) - 1/2\sqrt{(1 + r/\ell)^2 - 4/\ell}\,] \qquad (20)$$

and let

$$\sigma_1 = 1/2(1 - r/\ell) + 1/2\sqrt{(1 + r/\ell)^2 - 4/\ell} \qquad (21)$$

$$\sigma_2 = 1/2(1 - r/\ell) - 1/2\sqrt{(1 + r/\ell)^2 - 4/\ell} \qquad (22)$$

For $\sigma_1 \leq \sigma \leq \sigma_2$, $F(\sigma)$ is negative. For σ smaller than both σ_1 and σ_2, or greater than both σ_1 and σ_2, $\omega_I^2 \geq 0$. Next consider the locations of σ_1 and σ_2; there are two cases:

(a) If $r \geq \ell$, then $\sigma_2 < 0$ or complex
  (i) If $r > 1$, then $1 > \sigma_1 > 0$
  (ii) If $r = 1$, then $\sigma_1 = 0$
  (iii) If $r < 1$, then $\sigma_1 < 0$ or complex
(b) If $r < \ell$, then $\sigma_2 < 0$ or complex
  (i) If $r > 1$, then $\sigma_2 < 0$ or complex
  (ii) If $r = 1$, then $\sigma_2 = 0$
  (iii) If $r < 1$, then $1 > \sigma_2 > 0$ or complex

In Table 2, the first two columns correspond to different relative values of r and ℓ. The third column contains the regions in the first quadrant of the complex frequency plane for which ω^2 is less than or equal to ω_I^2 and for which ω_I^2 is non-negative. The region in the fourth quadrant is found by symmetry. The fourth column contains the active regions on the axes of the complex frequency plane found by taking the intersection of the regions in the second column with the axes. In the table the maximum real active frequency is denoted by ω_i.

VI. Region of Non-Negative ω_{II}^2 and Active Region for $r \geq \ell$.

Rewriting (15), we have

$$[\sigma - (1 - 1/r)]^2 + \omega_{II}^2 = [1/r(1/r - 1)] \qquad (23)$$

We see that ω_{II}^2 describes a circle with center at $(1 - 1/r, 0)$ and radius $R = \sqrt{1/r(1/r - 1)}$ as in Fig. 3.

If $r = 1$, the region inside the circle degenerates to a single point (0,0). σ_j, the maximum value in the region on the real axis, cannot exceed

1/2. For $r \geq \ell$ ω_{II}^2 is dominant. Combining the results for the axis from Table 1, with the dominant region for ω_{II}^2, we then have the active regions in Table 3 for $r \geq \ell$.

VII. Region of Non-Negative ω_{III}^2 and Active Region for $r < \ell$.

Since ω_{II}^2 is dominant, only for $r < \ell$ and $1/2(1-r/\ell) \leq \sigma \leq 1$, we need examine ω_{II}^2 only for this range of r, ℓ, and σ. Consider (16) rewritten as

$$\omega_{III}^2 = \frac{1}{2\ell\sigma+r}\left[1-(2\ell\sigma+r)(\sigma-1)^2\right] \tag{24}$$

Since ℓ, r and σ are positive, $\frac{1}{2\ell\sigma+r}$ is always positive.

Define $G(\sigma)$ by the equation

$$G(\sigma) = 1-(2\ell\sigma+r)(\sigma-1)^2 \tag{25}$$

Differentiating $G(\sigma)$ with respect to σ and setting the result equal to zero, we find that the minimum values of $G(\sigma)$ occur at $\sigma=1/3(1-r/\ell)$ and 1. We neglect the value at $\sigma=1$, since it is not in the range 0 to $1/2(1-r/\ell)$. We then have for the minimum value of $G(\sigma)$,

$$G(\sigma)_{\text{Min}} = 1 - 1/27\,(2\ell + r)(2 + r/\ell)^2 \tag{26}$$

The different regions of non-negative ω_{III}^2 are then easily found for different values of $G_{\text{Min}}(\sigma)$. We see that for values of r and ℓ satisfying the constraint $3\ell^{2/3} - 2\ell < r < 2\sqrt{\ell} - \ell$, there are two separate regions on the real axis which degenerate to one region for $r \leq 3\ell^{2/3} - 2\ell$ or $r \geq 2\sqrt{\ell} - \ell$.

Having now examined ω_{I}^2 in Section V, ω_{II}^2 in Section VI, and ω_{III}^2 in this section, we can describe the active regions for $r < \ell$. On the real and imaginary axes the active regions are as given in Table 1. Off the real axis ω_{III}^2 is dominant for $0 \leq \sigma < 1/2(1-r/\ell)$ and ω_{II}^2 is dominant for $1/2(1-r/\ell) \leq \sigma \leq 1$. The active regions for $r < \ell$ are plotted in Table 4. In its dominant interval ω_{II}^2 has its maximum value of $\omega_b^2 = 1/\ell - 1/4(1+r/\ell)^2$ at $\sigma=1/2(1-r/\ell)$. ω_b is positive so long as long as $r > 2\sqrt{\ell}-\ell$. If $r=2\sqrt{\ell}-\ell$, $\omega_b = 0$.

VIII. Synthesis of Passive Imbedding Networks.

As shown by Newcomb,[4] simple RLC imbedding networks can be devised to imbed a given active one-port network. One such imbedding network is included here with its justification. Consider as a general imbedding network a series LC circuit with impedance,

$$Z_p(p_0)=Lp_0 + \frac{1}{p_0 C} \tag{27}$$

$$= \sigma_0 L + \frac{\sigma_0}{\sigma_0^2 + \omega_0^2}\, 1/C + j\,[\,\omega_0 L - \frac{\omega_0}{\sigma_0^2 + \omega_0^2}\, 1/C] \tag{28}$$

The sufficient condition for N with impedance $Z(p_0)$ to be active at p_0 is that $Z_p(p_0) + Z(p_0) = 0$. We therefore require that

$$Z_p(p_0) = -Z(p_0) = -\alpha - j\beta \tag{29}$$

where $Z(p_0)$ is the impedance of the tunnel diode. Equating (28) and (29), we have

$$\sigma_0\ L + \frac{\sigma_0}{\sigma_0^2 + \omega_0^2}\, 1/C = -\alpha \tag{30a}$$

$$\omega_0 L - \frac{\omega_0}{\sigma_0^2 + \omega_0^2}\, 1/C = -\beta \tag{30b}$$

Solving for L and C, we have

$$L = \frac{-(\alpha\ \omega_0 + \beta\omega_0)}{2\sigma_0\ \omega_0} \tag{31a}$$

$$C = \frac{2\sigma_0 \omega_0}{(\beta\sigma_0 - \alpha\omega_0)(\sigma_0^2 + \omega_0^2)} \tag{31b}$$

In order that L and C be positive and real (note that ω_0 is assumed positive), it is necessary that

$$\alpha\ \omega_0 + \beta\sigma_0 \leq 0 \tag{32a}$$

$$\beta\sigma_0 - \alpha\omega_0 \geq 0 \tag{32b}$$

Equation (32a) corresponds to (5a) and (32b) corresponds to (5b). Hence if p_0 as an active point, then L and C are positive and real. The element values are given by

$$L = -\frac{1}{2\sigma_0\ \omega_0}\ [\ (r + \ell\ \sigma_0 + \frac{\sigma_0 - 1}{(\sigma_0 - 1)^2 + \omega_0^2})\omega_0 + \sigma_0\,(\ell\omega_0 - \frac{\omega_0}{(\sigma_0 - 1)^2 + \omega_0^2})\ ]$$

$$C = \frac{2\sigma_0\omega_0}{(\sigma_0^2 + \omega_0^2)[(\ell\omega_0 - \frac{\omega_0}{(\sigma_0-1)^2+\omega_0^2})\sigma_0 - \omega_0(r+\ell\sigma_0 + \frac{\sigma_0-1}{(\sigma_0-1)^2+\omega_0^2})]} \tag{33b}$$

Thus the series LC circuit covers the whole active range. As an example, consider a tunnel diode with $r=0.01$, $\ell=0.085$, and $p_0 = 0.1 + j0.1$. Then $L = 55.0$ and $C=10.0$. The resulting network is given in Fig. 4.

IX. Conclusion.

We have determined the region of activity in the right half of the complex frequency plane for the complete linear equivalent circuit of the tunnel diode. The active regions are described in Tables 2, 3 and 5 for various relative values of bulk resistance and lead inductance. Several interesting general conclusions can be drawn from the tables. As r and ℓ increase, the region of activity generally becomes smaller. For $r < \ell$ as r increases from zero, the active region first consists of a single domain. The region off the real axis then splits into two parts. One of these regions then disappears; the region on the real axis then splits, and finally for $r=1$ all that is left is one part of the real axis. The behavior for $r \geq \ell$ is not quite as interesting but still points out that as r increases the region of activity is severely limited.

Golosman and Newcomb[2] have pointed out that for two identical uncoupled tunnel diodes the region of activity is larger than for a single diode and is given by Column 3 in Table 2. The problem of determining the region of activity for multiple, non-identical tunnel diodes remains an unsolved problem.

REFERENCES

1. B. K. Kinariwala, "Theoretical Limitations on the Esaki Diode as a Network Element," IRE NEREM Record, 2, pp. 86-87; 1960.

2. B. S. Golosman and R. W. Newcomb, "Natural Frequencies of Tunnel Diodes with Parasitics," Tech. Report No. 2250-2, Stanford Electronics Laboratories, Stanford University, May 1962.

3. C. A. Desoer and E. S. Kuh, "Bounds on Natural Frequencies of Linear Active Networks," Proc. Symp. on Active Networks and Feedback Systems, pp. 415-436, Polytechnic Institute of Brooklyn; 1960.

4. R. W. Newcomb, "Synthesis of Passive Networks for Networks Active at P_0 - I," *1961 IRE Intl. Conv. Record*, part 4, pp. 162-175.

5. G. Birkhoff and S. MacLane, *A Survey of Modern Algebra*, The MacMillan Co., New York, 1953.

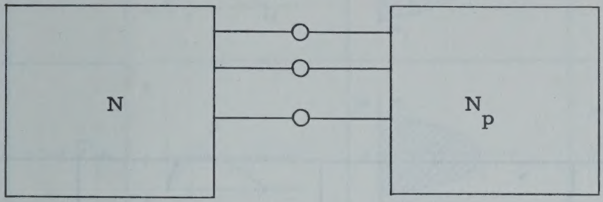
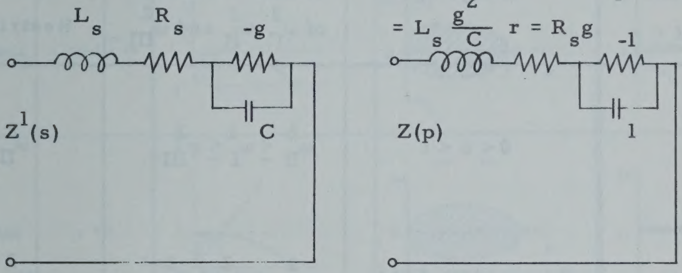


Fig. 1 Active Network N Imbedded in Passive Network N_p


L_s
R_s
$-g$
$Z^1(s)$
C
$= L_s \frac{g^2}{C}$
$r = R_s g$
-1
$Z(p)$
1

(a) Unnormalized

(b) Normalized

Fig. 2 Equivalent Circuit of Tunnel Diode

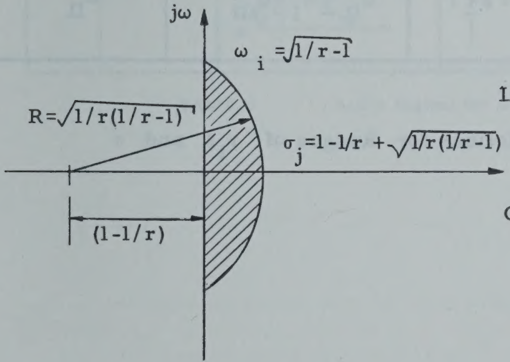

$j\omega$
$\omega_i = \sqrt{1/r - 1}$
$R = \sqrt{1/r(1/r - 1)}$
$\sigma_j = 1 - 1/r + \sqrt{1/r(1/r - 1)}$
σ
$(1 - 1/r)$

Fig. 3 Region in Right Half Plane with $\omega_{II}^2 \geq 0$ and $\omega^2 \leq \omega_{II}^2$

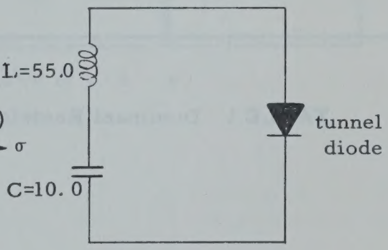

L=55.0
C=10.0
tunnel
diode

Fig. 4 Tunnel Diode Imbedded In Passive Network

Relative Values of r and ℓ	Range of σ	Relative Values of ω_I^2, ω_{II}^2 and ω_{III}^2	Dominant Restriction
$r \geq \ell$	$0 \leq \sigma \leq 1$	$\omega_{II}^2 \leq \omega_I^2 \leq \omega_{III}^2$	ω_{II}^2
$r < \ell$	$0 \leq \sigma < 1/2(1-r/\ell)$	$\omega_{III}^2 \leq \omega_I^2 \leq \omega_{II}^2$	ω_{III}^2
	$\sigma = 1/2\ (1-r/\ell)$	$\omega_I^2 = \omega_{II}^2 = \omega_{III}^2$	$\omega_I^2 = \omega_{II}^2 = \omega_{III}^2$
	$1/2(1-r/\ell) < \sigma \leq 1$	$\omega_{II}^2 \leq \omega_I^2 \leq \omega_{III}^2$	ω_{II}^2

TABLE 1 Dominant Restrictions for Ranges of $r_1 \ell$ and σ

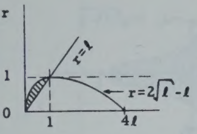
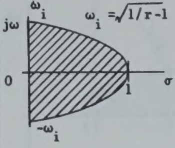
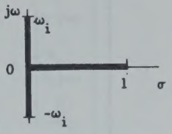
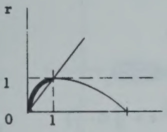
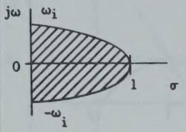
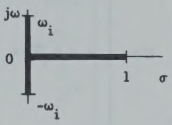
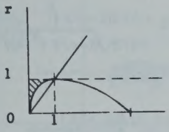
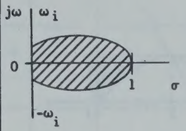
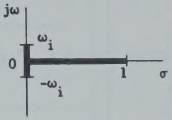
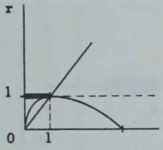
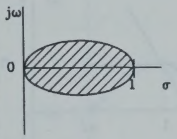
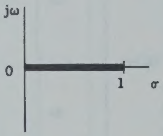
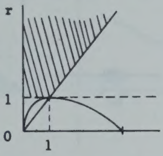
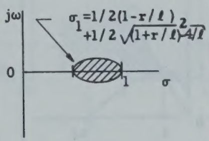
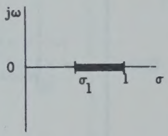
Case	Restrictions on r and ℓ		Region Enclosed by ω_I for $\omega_I^2 \geq 0$	Active Regions on Axes
I	$r < 2\sqrt{\ell}-\ell$	 r $r=\ell$ 1 $r=2\sqrt{\ell}-\ell$ 0 1 4ℓ	 ω_i $\omega_i=\sqrt{1/r-1}$ jω 0 1 σ $-\omega_i$	 jω ω_i 0 1 σ $-\omega_i$
IIa	$r = 2\sqrt{\ell}-\ell$		 jω ω_i 0 1 σ $-\omega_i$	 jω ω_i 0 1 σ $-\omega_i$
IIIaiii	$1 > r > 2\sqrt{\ell}-\ell$		 jω ω_i 0 1 σ $-\omega_i$	 jω ω_i 0 1 σ $-\omega_i$
IIIaii	$r = 1$			
IIIai	$r > 1$		 jω $\sigma_1=1/2(1-r/\ell)+1/2\sqrt{(1+r/\ell)^2-4/\ell}$ 0 1 σ	 jω 0 σ_1 1 σ

TABLE 2 Active Regions for $\omega = 0$ or $\sigma = 0$, $r \geq \ell$

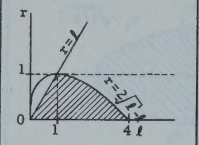
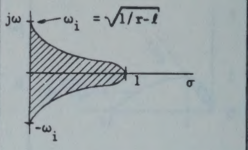
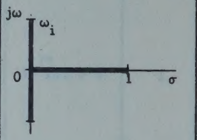
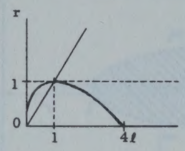
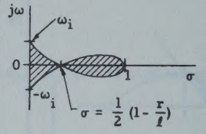
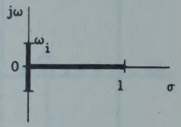
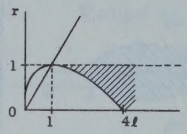
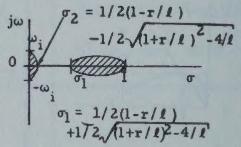
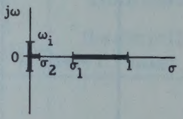
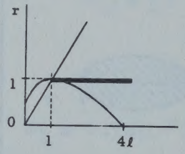
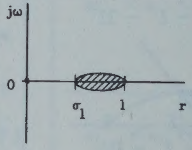
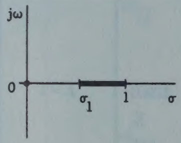
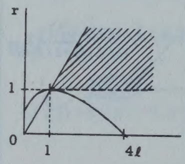
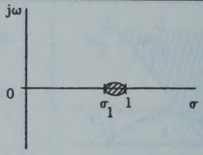
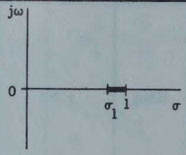
Case	Restrictions on r and ℓ		Region Enclosed by ω_I for $\omega_I^2 \geq 0$	Active Regions on Axes
I	$r < 2\sqrt{\ell} - \ell$	 r $r=\ell$ 1 $r=2\sqrt{\ell}-\ell$ 0 1 4 ℓ	 jω $\omega_i = \sqrt{1/r-\ell}$ 1 σ $-\omega_i$	 jω ω_i 0 1 σ
IIb	$r = 2\sqrt{\ell} - \ell$		 jω ω_i 0 1 σ $-\omega_i$ $\sigma = \frac{1}{2}(1-\frac{r}{\ell})$	 jω ω_i 0 1 σ
IIIbiii	$1 > r > 2\sqrt{\ell} - \ell$		 jω $\sigma_2 = 1/2(1-r/\ell) -1/2\sqrt{(1+r/\ell)^2-4/\ell}$ ω_i 0 σ_1 1 σ $-\omega_i$ $\sigma_1 = 1/2(1-r/\ell) +1/2\sqrt{(1+r/\ell)^2-4/\ell}$	 jω ω_i 0 σ_2 σ_1 1 σ
IIIbii	$r = 1$		 jω 0 σ_1 1 r	 jω 0 σ_1 1 σ
IIIbi	$r > 1$		 jω 0 σ_1 1 σ	 jω 0 σ_1 1 σ

TABLE 2 Active Regions for $\omega = 0$ or $\sigma = 0$, $r \geq \ell$ (cont.)

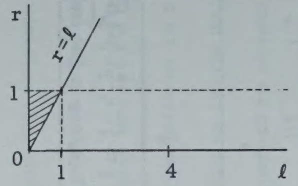
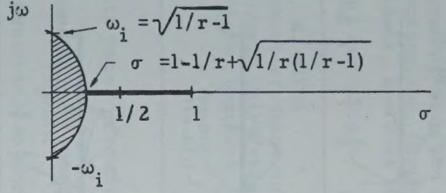
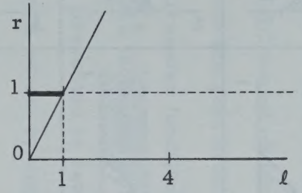
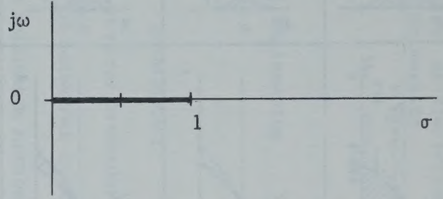
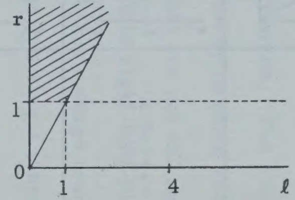
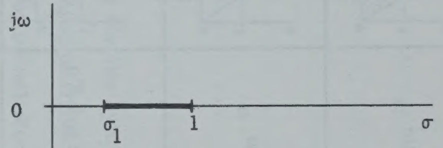
Constraints on r and ℓ		Active Regions
i $r < 1$		 jω $\omega_i = \sqrt{1/r-1}$ $\sigma = 1-1/r+\sqrt{1/r(1/r-1)}$ 1/2 1 σ $-\omega_i$
ii $r = 1$		
iii $r > 1$		 jω 0 σ_1 1 σ $\sigma_1 = \frac{1}{2}(1-\frac{r}{\ell})+\frac{1}{2}\sqrt{(1+\frac{r}{\ell})^2-\frac{4}{\ell}}$

TABLE 3 Active Regions for $r \geq \ell$

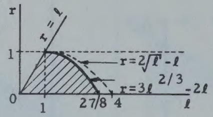
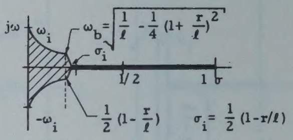
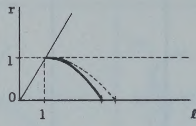
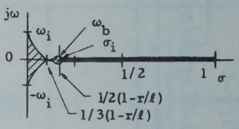
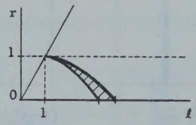
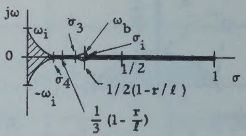
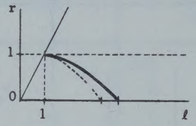
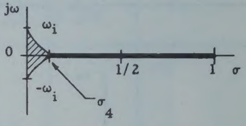
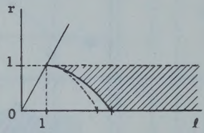
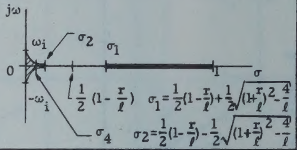
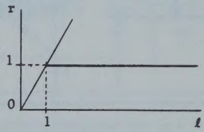
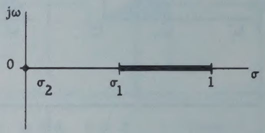
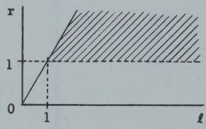
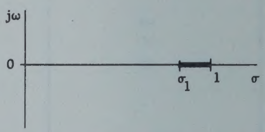
Constraints on r and ℓ		Active Regions
i $r < 3\ell^{2/3} - 2\ell$	 r $r = \ell$ 1 0 1 27/8 4 ℓ $r = 2\sqrt{\ell} - \ell$ $r = 3\ell^{2/3} - 2\ell$	 jω ω_i $\omega_b = \sqrt{\frac{1}{\ell} - \frac{1}{4}(1 + \frac{r}{\ell})^2}$ σ_i 1/2 1 σ $-\omega_i$ $\frac{1}{2}(1 - \frac{r}{\ell})$ $\sigma_i = \frac{1}{2}(1 - r/\ell)$
ii $r = 3\ell^{2/3} - 2\ell$		 jω ω_i ω_b σ_i 0 1/2 1 σ $-\omega_i$ $1/2(1 - r/\ell)$ $1/3(1 - r/\ell)$
iii $3\ell^{2/3} - 2 < r < 2\sqrt{\ell} - \ell$		 jω ω_i σ_3 ω_b σ_i 0 1/2 1 σ σ_4 $-\omega_i$ $1/2(1 - r/\ell)$ $\frac{1}{3}(1 - \frac{r}{\ell})$
iv $r = 2\sqrt{\ell} - \ell$		 jω ω_i 0 1/2 1 σ $-\omega_i$ σ_4
v $1 > r > 2\sqrt{\ell} - \ell$		 jω ω_i σ_2 σ_1 0 1 σ $-\omega_i$ $\frac{1}{2}(1 - \frac{r}{\ell})$ $\sigma_1 = \frac{1}{2}(1 - \frac{r}{\ell}) + \frac{1}{2}\sqrt{(1 + \frac{r}{\ell})^2 - \frac{4}{\ell}}$ σ_4 $\sigma_2 = \frac{1}{2}(1 - \frac{r}{\ell}) - \frac{1}{2}\sqrt{(1 + \frac{r}{\ell})^2 - \frac{4}{\ell}}$
vi $r = 1$		 jω 0 σ_2 σ_1 1 σ
vii $r > 1$		 jω 0 σ_1 1 σ

TABLE 4 $G(\sigma)$ vs σ Heavy Lines Indicate Region of σ Where $G(\sigma) \geq 0$ for $0 \leq \sigma \leq \sigma_i$

TRANSFER FUNCTION SYNTHESIS OF RC-TUNNEL-DIODE NETWORKS*

D. L. LOSEE
Department of Physics, University of Illinois
Urbana, Illinois
Formerly with the School of Electrical
Engineering, Cornell University
Ithaca, New York

and

S. K. MITRA
School of Electrical Engineering
Cornell University
Ithaca, New York

SUMMARY

This paper presents several procedures for the synthesis of various transfer functions by means of networks containing positive resistances, positive capacitances, and a minimum number of tunnel diodes (represented by a parallel combination of a positive capacitor and a negative resistor). Use of tunnel diodes are shown to allow realization of arbitrary transmission zeroes by means of grounded RC tunnel-diode networks. The tunnel diode may be employed either to improve the gain obtainable by passive realization techniques or to synthesize transfer functions that are not RC realizable. In addition to a cascade synthesis technique to realize right-half-plane transmission zeroes, a method is presented to realize any voltage transfer ratio with unrestricted transmission zeroes by employing a single tunnel diode.

I. INTRODUCTION

The study of networks obtained by imbedding tunnel diodes in a passive structure has received considerable attention in recent literature[1-7], and the references cited are just a few out of a much longer list. In the area of transfer-function synthesis of networks containing resistances, capacitances and tunnel diodes, Weinberg[1] has presented a general method of synthesis by means of a "reverse predistortion" technique. In his approach, the number of tunnel diodes required is nearly equal to the number of capaci-

---

* This work was supported by the Air Force Avionics Laboratory, Air Force Systems Command, WPAFB under Contract No. AF 33(657)-10682. Further reproduction is authorized to satisfy the needs of U. S. Government.

tances in the network. The purpose of this paper is to present some straightforward procedures for the realization of transfer functions by means of networks containing resistances, capacitances and a minimum number of tunnel diodes. For convenience, such networks are denoted by $RC{:}t_d^{(N)}$ networks, where the superscript N refers to the number of tunnel diodes in the network. The model of a tunnel diode (as shown in Figure 1), used by most of the authors, is chosen for discussion. This paper is concerned only with the a-c circuit of the diode, and no consideration is given to the problem of biasing, the dynamic range of operation and other related problems. Though, a $RC{:}t_d$ transfer function can have zeroes anywhere in the complex plane and poles only on the real axis, we restrict our attention to stable transfer functions only.

In the next three sections, we discuss several new methods for the realization of $RC{:}t_d$ transfer functions. The methods presented stem directly from standard passive synthesis techniques and yield networks with a minimum number of tunnel diodes. Sections II and III deal with transfer functions having transmission zeroes on the real axis and in the left-half plane, respectively. In Section IV, we consider the more general case of arbitrary transmission zeroes. We show in Section IV. B, the sufficiency of a single tunnel diode, in synthesizing a transfer function with unrestricted transmission zeroes.*

II. TRANSMISSION ZEROES ON THE REAL AXIS

A. NON-MINIMUM PHASE TRANSFER FUNCTIONS

Perhaps the simplest of all RC transfer function synthesis techniques is the procedure of zero shifting. This method permits the realization of negative real transmission zeroes by means of a ladder network. The introduction of a tunnel diode as a permissible branch makes the realization of positive real zeroes possible. The procedure, a direct extension of the familiar zero-shifting process, has been independently presented by Mitra[4] and Su[6]. Briefly, the technique is as follows:

---

*A very restricted class of such transfer functions (when the degree of the numerator and denominator polynomials are equal) has been discussed by Herskowitz and Ghausi[5].

Given a RC:t_d open-circuit transfer impedance $z_{21}(s)$, a passive RC driving-point impedance $z_{11}(s)$ having the same poles as $z_{21}(s)$ is selected. Next, $z_{11}(s)$ is developed into a ladder network in the form of Figure 2, where the shunt resistances R_i are chosen to create a zero of the remainder admittance $Y_i(s)$ at the desired location on the positive real axis. This zero is removed as a parallel combination of negative resistance and a positive capacitance in the series branch. The remaining impedance $Z_j(s)$ is guaranteed to be a passive RC impedance. There is no specified order for removing the positive real axis zeroes, but the order of removal influences the capacitance and the conductance values of the tunnel diodes. As a result, one order may be preferable to others with respect to diodes desired. In this method of realization, the number of tunnel diodes necessary is equal to the number of positive real axis zeroes.

B. MINIMUM-PHASE TRANSFER FUNCTIONS

The use of tunnel diodes in RC circuits is not strictly confined to synthesis problems which could not be solved using passive elements alone, but the tunnel diode may be usefully employed to improve upon the results obtainable through passive synthesis techniques.

Specifically, when it is required that the network design include an internal resistance in the source driving the network, the tunnel diode may be used to "compensate" for the required internal resistance. To see this, consider the realization of the transfer voltage ratio,

$$t_v = \frac{E_2}{E_1} = H \frac{\prod_{i=1}^{m} (s+\sigma_i)}{\prod_{j=1}^{n} (s+z_j)} \quad , \quad m \leq (n-1) \quad , \tag{1}$$

and suppose it is desired that the realization include a one-ohm series resistance at the input of the network.

To realize this transfer ratio, let us select an RC:$t_d^{(1)}$ driving-point impedance,[7] $z_{11}(s)$ having zeroes given by the denominator of t_v,

$$z_{11}(s) = \frac{\prod_{j=1}^{n} (s + z_j)}{(s-\alpha) \prod_{k=1}^{n-1} (s + p_k)} , \tag{2}$$

where α, z_j and p_k are real and positive. Now it will be shown that $z_{11}(s)$ is always realizable in the form indicated in Figure 3, where the RC ladder realizes the finite transmission zeroes of t_v. First, we remove the one-ohm series resistance from $z_{11}(s)$. This leaves a remainder function having all negative real zeroes and all but one negative real poles. Then the remainder admittance has the form

$$Y_o(s) = \frac{K(s - \alpha) \prod_{i=1}^{n-1} (s + p_i)}{\prod_{j=1}^{n-1} (s + \beta_j)} = \frac{1}{z_{11}(s) - 1} , \tag{3}$$

where $\beta_i < p_i$. Now $Y_o(s)$ can be written as

$$Y_o(s) = s\, C_d - G_d + Y_r(s) , \tag{4}$$

where $Y_r(s)$ is a passive RC admittance and

$$C_d = \frac{1}{\sum_{i=1}^{n} z_i - \sum_{k=1}^{n-1} p_k + \alpha} , \tag{5}$$

$$G_d = \frac{\alpha \prod_{k=1}^{n-1} p_k}{\prod_{i=1}^{n} z_i + \alpha \prod_{k=1}^{n-1} p_k} . \tag{6}$$

The validity of Equation (4) can be easily checked by making a partial fraction expansion of $Y_o(s)/s$. It should be noted that C_d will create transmission zeroes at infinity, thus the procedure is valid only if t_v has at least one transmission zero at infinity. The remainder of the zeroes $\{\sigma_i\}$, are obtained by using zero-shifting technique in the development of $Y_r(s)$.

By employing a tunnel diode to realize a passively realizable transfer function, the possible gain constant H in Equation (1) is raised.* This will be clear from a special case when all the transmission zeroes are at infinity. In this case, the network will be of the form shown in Figure 4. From Figure 4 and Equation (1), we see that

$$H = a \prod_{j=1}^{n-1} p_j + \prod_{i=1}^{n} z_i \quad . \tag{7}$$

Since the maximum value of H obtainable by a grounded RC network is $\prod_{i=1}^{n} z_i$, we see that H obtained with a $RC{:}t_d^{(1)}$ two-port is greater.

Example 1: Consider

$$t_v = \frac{H}{(s+1)(s+3)} \quad .$$

Choosing

$$z_{11} = \frac{(s+1)(s+3)}{(s-2)(s+2)} \quad ,$$

we obtain the network of Figure 5 for the realization of the above voltage transfer ratio. The value of H obtained is 7, while that obtainable by a passive network is 3.

III. TRANSMISSION ZEROES IN THE LEFT HALF PLANE

In this section, it will be shown that we can realize, by means of a $RC{:}t_d^{(1)}$ two-port, any $RC{:}t_d$ transfer voltage ratio t_v, whose zeroes are in the left half plane, with the restriction that there be at least one zero at infinity. Consider the realization of a transfer voltage ratio $t_v = E_2/E_1$, in the form shown in Figure 6. In terms of the y-parameters of the RC two-port, we have

---

* An approximate bound has been derived by Losee[8] on the amount of improvement possible for H.

$$t_v = \frac{E_2}{E_1} = \frac{-y_{21}}{y_{22}-1} \quad . \tag{8}$$

Given a transfer voltage ratio as a ratio of two polynomials, N(s)/D(s), the problem is then to express t_v as,

$$t_v = \frac{N(s)}{D(s)} = \frac{N(s)}{D_1(s) - D_2(s)} = \frac{N(s)/D_2(s)}{\dfrac{D_1(s)}{D_2(s)} - 1} \quad , \tag{9}$$

and identify

$$y_{22} = \frac{D_1(s)}{D_2(s)} \quad ; \quad -y_{21} = \frac{N(s)}{D_2(s)} \quad , \tag{10}$$

in such a way so that y_{22} and $-y_{21}$, as given by Equation (10), form a realizable pair for the synthesis of the RC two-port. Furthermore, if y_{22} has a pole at infinity (this implies t_v should have at least a transmission zero at infinity), a shunt capacitance across the negative resistance can be produced as the first step of the realization procedure. The combination of the positive capacitance in parallel with the negative resistance can finally be replaced by a tunnel diode. The main problem can be stated:

Given a polynomial D(s), all of whose zeroes are negative real and simple, decompose it as

$$D(s) = D_1(s) - D_2(s) \tag{11}$$

where $D_1(s)/D_2(s)$ is an RC admittance with a pole at infinity. Let us choose a polynomial $D_2(s)$, of one degree less than D(s) and all of whose zeroes are negative real and alternating with the zeroes of D(s). Furthermore, the zero closest to the origin belongs to D(s). We see, from the plot of $D(\sigma)$ and $-D_2(\sigma)$ in Figure 7, that by properly choosing the leading coefficient of $D_2(s)$, a polynomial $D_1(s) = D(s) + D_2(s)$ can always be found, such that $D_1(s)/D_2(s)$ is an RC admittance with a pole at infinity. The following example illustrates the synthesis procedure:

Example 2: Let the specified t_v be

$$t_v = \frac{N(s)}{D(s)} = \frac{H}{s^2 + 3s + 1} = \frac{H}{(s + 0.382)(s + 2.618)} .$$

Choosing $D_2(s) = s + 2$, we obtain $D_1(s)$ as

$$D_1(s) = (s^2 + 3s + 1) + (s + 2) = s^2 + 4s + 3 = (s + 1)(s + 3) .$$

Hence,

$$y_{22} = \frac{(s + 1)(s + 3)}{(s + 2)} \quad ; \quad -y_{21} = \frac{H}{s + 2} .$$

The network realization is shown in Figure 8. The realized value of the multiplier constant H is found to be equal to 3, whereas the maximum H obtainable by passive RC two-port is 1.

IV. UNRESTRICTED TRANSMISSION ZEROES

A. CASCADE SYNTHESIS TECHNIQUE

Dasher's method[9] of realizing a given set of complex transmission zeroes in the left half plane, including the $j\omega$-axis* along with a specified driving-point function by means of cascaded, "canonic" zero-producing sections, can also be applied to realize transmission zeroes in the right half plane if active elements are used in the zero producing sections. For example, the canonic two-port shown in Figure 9 can realize either a conjugate pair or a pair of real zeroes in the right half plane† of the form

$$t(s) = s^2 - 2\alpha s + \beta^2 , \tag{12}$$

---

* Complex zero-producing sections have been suggested by Hakimi and Seshu[12] to realize transmission zeroes in the right half plane.

† Recently, Su[6] has extended Dasher's technique to include right half plane zeroes using "canonic" sections which involve only a negative resistance without a shunt capacitance.

where α and β are real and positive. It should be noted that a single positive real axis zero can be realized by the zero-shifting technique outlined in Section II. A.

The realization procedure described is identical to Guillemin's treatment of Dasher's technique.** First, the given RC driving-point function is expressed in a partial fraction expansion

$$Z_1(s) = \rho_\infty + \frac{\rho_o}{s} + \sum_{i=1}^{n} \frac{k_i}{(s - s_i)} \quad . \tag{13}$$

Next, we define

$$m_i = \frac{(-s_i)k_i}{t^2(+s_i)} \quad , \tag{14}$$

and

$$s_a = \frac{\sum m_i s_i}{\sum m_i} \quad . \tag{15}$$

It can be shown that by proper "preparation" of Z_1, we can find a positive real value for K satisfying the following two inequalities:

$$0 < K - \beta^2 \sum_{i=1}^{n} \frac{m_i}{-s_i}(s_i - s_a) \leqq \frac{(-s_a)\rho_o}{\beta^2} \quad , \tag{16}$$

and

$$0 < K - \sum_{i=1}^{n} m_i s_i (s_i - s_a) \leqq \rho_\infty \tag{17}$$

The open-circuit impedance parameters of the zero-producing section can be made to have the following form:

$$z_{12} = K + \frac{k_o}{s} - \frac{k_{12}}{s - s_a} = K \frac{(s^2 - 2\alpha s + \beta^2)}{s(s - s_a)} \quad ,$$

** For details, see Guillemin.[10]

$$z_{11} = K + \frac{k_o}{s} + \frac{k_{12}/\mu}{s - s_a} , \tag{18}$$

$$z_{22} = K + \frac{k_o}{s} + \frac{\mu k_{12}}{s - s_a} .$$

For the extraction of the zero-producing section, we write

$$\frac{h(s)}{q(s)} = \frac{s(s - s_a)(z_{11} - Z_1)}{K^2 t^2(s)} = \frac{1}{K^2} \sum_{i=1}^{n} m_i \frac{(s_i - s_a)}{(s - s_i)} , \tag{19}$$

$$z_{11} - Z_1 = \frac{K^2 t^2(s) h(s)}{s(s - s_a) q(s)} , \tag{20}$$

and

$$z_{22} + Z_2 = \frac{q(s)}{s(s - s_a) h(s)} . \tag{21}$$

The terminating impedance Z_2 is always an RC impedance and two degrees lower than the degree of Z_1. Finally, the element values of the zero-producing section are given as:

$$R_1 = \frac{K(1 + \mu)}{\mu} ,$$

$$R_2 = K(1 + \mu) ,$$

$$G_3 = \frac{\mu(-2\alpha s_a + \beta^2)}{K(1 + \mu)^2 t(s_a)} , \tag{22}$$

$$C_a = \frac{-\mu s_a}{K(1 + \mu)^2 t(s_a)} ,$$

and

$$C_b = \frac{-s_a}{K\beta^2} .$$

On the admittance basis, a dual procedure can similarly be formulated.

Example 3: Let us develop

$$Z(s) = \frac{(s + 1)(s + 3)}{(2s + 1)(s + 2)} \tag{23}$$

into a cascade network so that a transmission zeroes are given by

$$t(s) = s^2 - 2s + 2.$$

This example was solved on a digital computer using a program by Rothfarb,[12] and the final network is shown in Figure 10.

B. SYNTHESIS USING A SINGLE TUNNEL DIODE

In the following discussions, it will be shown that only one tunnel diode is necessary in the realization of any voltage transfer ratio with unrestricted transmission zeroes. Let

$$t_v = \left.\frac{E_2}{E_1}\right|_{I_2 = 0} = \frac{-y_{21}}{y_{22}} = \frac{N(s)}{D(s)} \tag{24}$$

be the specified transfer function. Consider the realization of t_v by means of a parallel connection of two two-ports (see Figure 11). If we write $\left\{y_{ij}^{(a)}\right\}$ and $\left\{y_{ij}^{(b)}\right\}$ as the y-parameters of the networks N_a and N_b, respectively, then

$$t_v = \frac{-y_{21}}{y_{22}} = \frac{-y_{21}^{(a)} - y_{21}^{(b)}}{y_{22}^{(a)} + y_{22}^{(b)}} \;. \tag{25}$$

Because of the assumed configuration, the transmission zeroes of N_b will be on the negative real axis and N_a will have one transmission zero on the positive real asis and the remaining on the negative real axis. As a result, we can write

$$-y_{21}^{(a)} = \frac{(s - a)\, n_{21}'(s)}{d(s)} \;, \qquad -y_{21}^{(b)} = \frac{n_{21}''(s)}{d(s)} \;, \tag{26}$$

where $n_{21}'(s)$ and $n_{21}''(s)$ have all negative real roots. This implies that the numerator N(s) of the given transfer function will be of the form

$$N(s) = (s - a)n_{21}'(s) + n_{21}''(s) \quad . \tag{27}$$

Hence the problem of realizing an arbitrary $RC{:}t_d$ voltage transfer ratio, in the form of Figure 11, reduces to the problem of expressing its numerator in the form of Equation (27). To this end, it is necessary to show that:

Any real polynomial N(s) with positive leading coefficient can always be decomposed in the form of Equation (27).

Let us divide N(s) by a polynomial Q(s) of one degree less and having all negative real roots. The partial fraction expansion of N(s)/D(s) will in general be of the form:

$$\frac{N(s)}{Q(s)} = k_{oo}^{+} s + \sum_i \frac{k_i^{+}}{s + \sigma_i} - \sum_j \frac{k_j^{-}}{s + \sigma_j} - G \quad , \tag{28}$$

where $\left\{k_i^{+}\right\}$ and $\left\{k_i^{-}\right\}$ are real positive constant and G is a real constant. Grouping the terms

$$Y(s) = \sum_i \frac{k_i^{+}}{s + \sigma_i} \quad ; \qquad Y^{*}(s) = k_{oo}^{+} s - \sum_j \frac{k_j^{-}}{s + \sigma_j} - G, \tag{29}$$

we note that Y(s) is a RL driving-point admittance, and as a result, has all its zeroes and poles on the negative real axis. But $Y^{*}(s)$ is the driving-point admittance of a network containing positive resistances, positive capacitances and a single negative resistance, and therefore can have at most one positive real zero.[4] The remaining zeroes and the poles will be on the negative real axis.

Note that the decomposition of N(s) in the form of Equation (27) is not unique. The following example will illustrate the procedure.

Example 4: Consider the transfer function

$$t_v = \frac{N(s)}{D(s)} = H \frac{(s - 1)(s - 3)}{(s + 1)(s + 3)} \quad . \tag{30}$$

Here $N(s) = s^2 - 4s + 3$. Choosing $Q(s) = s$, we obtain the decomposition of $N(s)$ as

$$N(s) = s(s - 4) + 3 \quad .$$

An RC driving-point admittance is formed having the same zeroes as given by the denominator of t_v,

$$y_{22} = \frac{(s + 1)(s + 3)}{(s + 2)} \quad .$$

Next, y_{22} is developed into two ladder networks, each having the following two transfer admittances,

$$- y_{21}^{(a)} = \frac{s(s - 4)}{s + 2}$$

and

$$- y_{21}^{(b)} = \frac{3}{s + 2} \quad ,$$

respectively, within a multiplier constant. After proper impedance level adjustments, the final network is shown in Figure 12.

V. CONCLUSION

We have demonstrated that a single tunnel diode embedded in a passive RC structure is sufficient to realize a voltage transfer ratio with arbitrary zero pattern and poles restricted to negative real axis. The final realization is in the form of two parallel ladders. A major practical disadvantage of this type of synthesis is the difficulty of proper alignment. In a parallel ladder form of realization, the transmission zeroes (which are important in the case of RC-tunnel-diode transfer functions) are all affected by the adjustment of a single element value. This problem does not arise in the cascade-synthesis technique, where the transfer function is realized by means of a cascade of component two-ports of which each produces one transmission zero or a pair of complex conjugate zeroes. In the latter method, the number of tunnel diodes required is nearly equal to the number of pairs of right half plane zeroes. In the case of transfer functions having transmission zeroes on the negative real axis or in the left half plane, the synthesis technique presented in Sections II. B and III, respectively, are preferable from the point of view of biasing. In each of this two methods, a single tunnel diode is needed.

It has been shown that several advantages exist for including a tunnel diode in an otherwise passive structure, for example: (1) realization of a transfer function with a multiplier constant that is larger than the maximum obtainable with a transformerless RLC network, and (2) realization of non-minimum phase transfer functions by means of grounded $RC{:}t_d$ two-ports.

The problem of sensitivity of transfer functions with respect to variations of the tunnel diode constants, which has been considered by Losee[8] for a special case, deserves further study.

VI. ACKNOWLEDGMENT

The authors would like to thank Professor N. DeClaris for useful discussions.

VII. REFERENCES

1. L. Weinberg, "Synthesis Using Tunnel Diodes and Masers," I. R. E. Trans. on Circuit Theory, CT-8, pp. 66-75, (March 1961).

2. B. K. Kinariwala, "The Esaki Diode as a Network Element," I. R. E. Trans. on Circuit Theory, CT-8, pp. 389-397, (December 1961).

3. I. W. Sandberg, "The Realizability of Multi-port Structures Obtained by Imbedding a Tunnel Diode in a Lossless Reciprocal Network," Bell Systems Tech. J., 41, pp. 857-876, (May 1962).

4. S. K. Mitra, "Synthesis of Active RC Networks," Res. Report 60-442, Electronics Res. Lab., Univ. of Calif., Berkeley, Calif., (March 1962).

5. G. J. Herskowitz and M. S. Ghausi, "Transfer Function Synthesis Employing One Tunnel Diode and a Passive RC Ladder Network," 1962 WESCON Convention Record,

6. K. L. Su, "Cascade Synthesis of RC Networks Using Negative Resistances," I. R. E. Trans. on Circuit Theory, CT-9, pp. 423-425, (December 1962).

7. S. K. Mitra, "The Realizability of Tunnel-Diode-RC Networks," J. Franklin Inst., 275, pp. 205-216, (March 1963).

8. D. L. Losee, "RC-Tunnel-Diode Transfer Function Synthesis," M. S. Thesis, Cornell Univ., Ithaca, N. Y., (June 1963).

9. B. J. Dasher, "Synthesis of RC Transfer Functions as Unbalanced Two-Terminal Pair Networks," I. R. E. Trans. on Circuit Theory, CT-1 pp. 20-34, (December 1952).

10. E. A. Guillemin, "Synthesis of Passive Networks," New York: John Wiley and Sons, 1957.

11. S. L. Hakimi and S. Seshu, "Realization of Complex Zeroes of Transmission by means of RC Networks, "Proc. Nat. Elec. Conf., 13, pp. 1013-1025, (1957).

12. B. Rothfarb, "Digital Computer Realization of Complex Transmission Zeroes of RC Transfer Functions," Unpublished Senior Project Report, Cornell Univ., (June 1963).

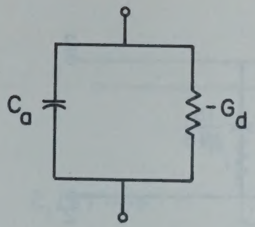


FIGURE 1. Equivalent Circuit of a Tunnel Diode.

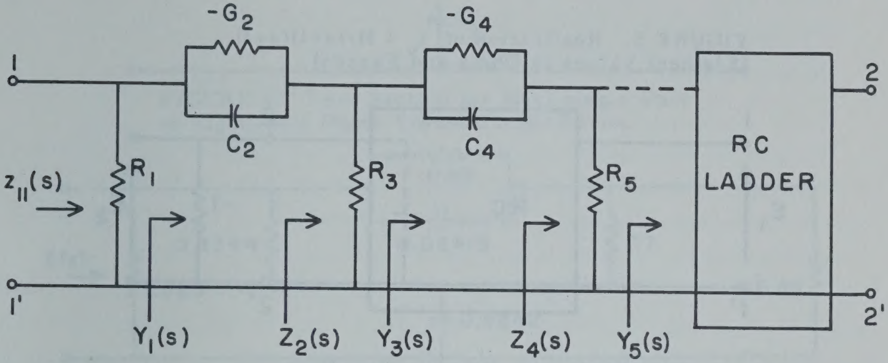

1
1'
2
2'
$-G_2$
$-G_4$
C_2
C_4
R_1
R_3
R_5
$z_{11}(s)$
$Y_1(s)$
$Z_2(s)$
$Y_3(s)$
$Z_4(s)$
$Y_5(s)$
RC LADDER

FIGURE 2. Realization of Real Axis Transmission Zeros.

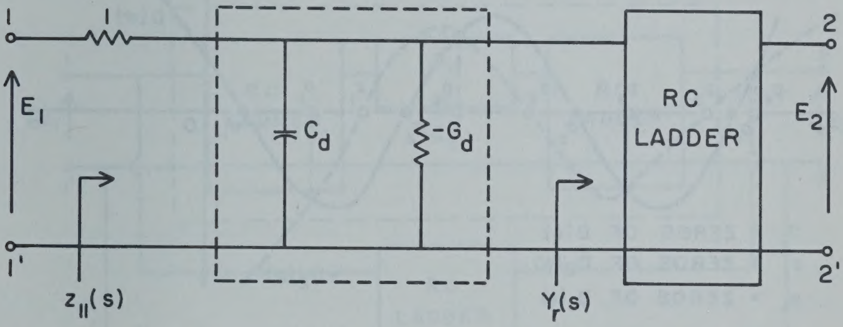

1
1'
2
2'
1
E_1
E_2
C_d
$-G_d$
RC LADDER
$z_{11}(s)$
$Y_r(s)$

FIGURE 3. Realization of Negative Real Axis Transmission Zeros.

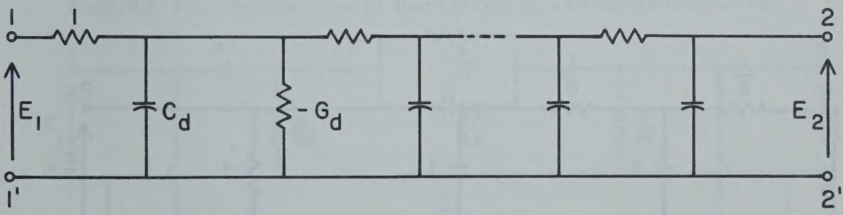

1
1'
2
2'
1
E_1
E_2
C_d
$-G_d$

FIGURE 4. Realization of All-Pole Transfer Voltage Ratio.

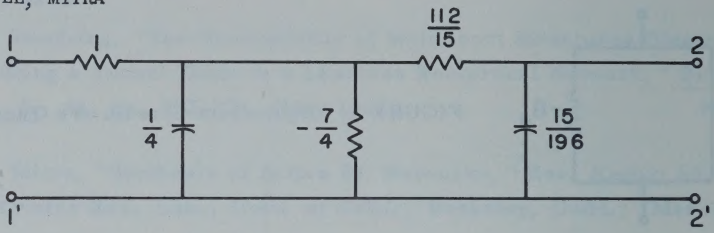

1
1
112/15
2
1/4
-7/4
15/196
1'
2'

FIGURE 5. Realization of $t_y = H/(s+1)(s+3)$, (Element Values in Ohms and Farads).

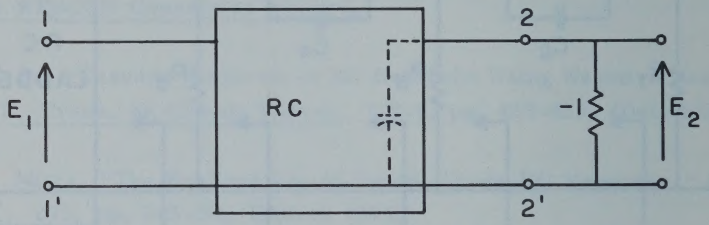

1
2
E₁
RC
-1
E₂
1'
2'

FIGURE 6. An $RC{:}t_d^{(1)}$ Two-Port.

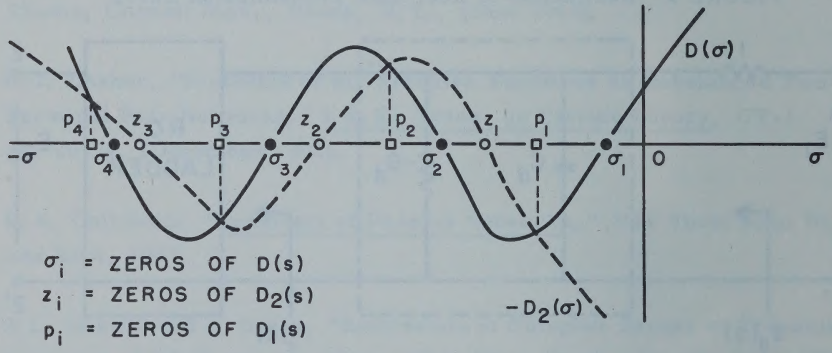


FIGURE 7. On the Decomposition of $D(s) = D_1(s) + D_2(s)$.

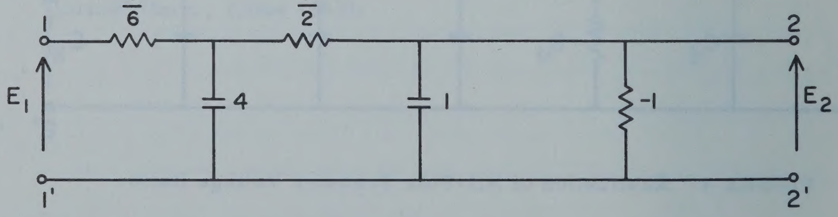

1
1/6
1/2
2
E₁
4
1
-1
E₂
1'
2'

FIGURE 8. Realization of $t_y = H/(s^2 + 3s + 1)$, (Element Values in Ohms and Farads).

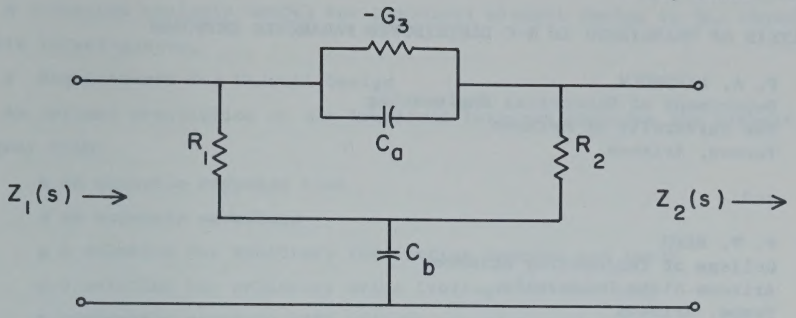

-G3
R1
Ca
R2
Z1(s) →
Z2(s) →
Cb

FIGURE 9. Zero Section for Realizing a Pair of Right-Half Plane Transmission Zeros.

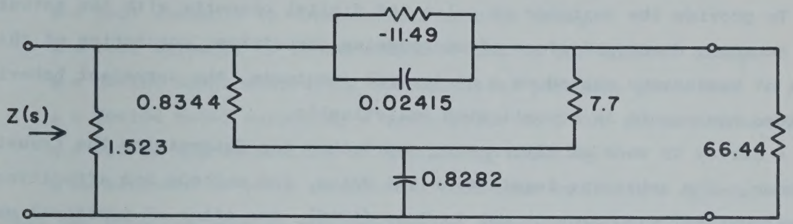

-11.49
0.8344
0.02415
7.7
Z(s)
1.523
66.44
0.8282

FIGURE 10. Example 3.

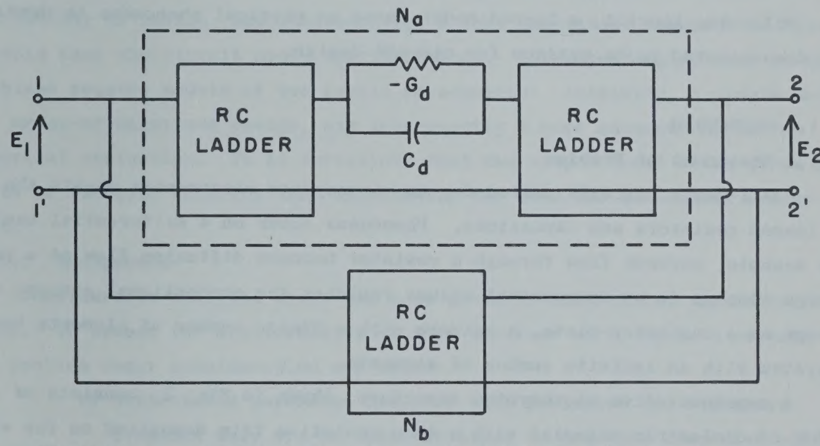

Na
1
2
E1
RC LADDER
-Gd
Cd
RC LADDER
E2
1'
2'
RC LADDER
Nb

FIGURE 11. Realization of Unrestricted Transmission Zeros.

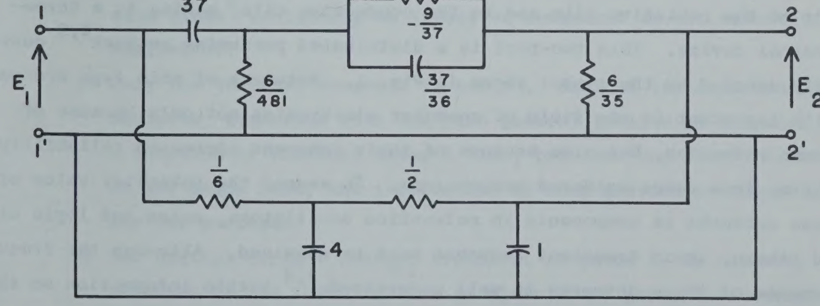

1
37
-9/37
2
E1
6/481
37/36
6/35
E2
1'
2'
1/6
1/2
4
1

FIGURE 12. Example 4.

ANALYSIS OF TRANSIENTS IN R-C DISTRIBUTED PARAMETER NETWORKS

F. A. LINDHOLM
Department of Electrical Engineering
The University of Arizona
Tucson, Arizona

and

W. W. HAPP
College of Engineering Sciences
Arizona State University
Tempe, Arizona

ABSTRACT

To provide the designer of pulse and digital circuits with the potential of a recently developed class of microsystem structures, consisting of thin films of resistance plated on a dielectric substrate, the transient behavior of these components is investigated analytically,

A survey is made of alternative approaches for determining the transient response. For arbitrary termination and drive, two methods are effective: one of mathematical significance due to Gray (1960), the other of practical use to the circuit designer pioneered by Linvill (1958).

Following Linvill, a lumped model based on physical phenomena is developed and demonstrated to be optimum for circuit design.

1. OBJECTIVE

1.1 Statement of Problem

Small-scale circuits called for in microsystem electronics negate the use of lumped resistors and capacitors. Phenomena occur on a differential basis; for example, current flow through a resistor becomes diffusion flow at a point, charge storage in an incremental volume replaces the conventional concept of charge on a capacitor plate, a network with a finite number of elements becomes a system with an infinite number of elements.

A representative microsystem structure, shown in Fig. 1, consists of a wafer of dielectric material with a thin resistive film deposited on top and a conductive film plated on the bottom. Electrical contracts are made to the ends of the resistive film and to the conductive film, making it a three-terminal device. This two-port is a distributed parameter network[1,2] conveniently denoted by the symbol shown in Fig. 1. Networks of this type are potentially important to the field of computer electronics not only because of volume reduction, but also because of their inherent increased reliability arising from fewer soldered connections. To assess the potential value of these networks as components in relaxation oscillators, pulse and logic circuits and others, their transient response must be examined. Although the frequency response of these networks is well understood,[3,4] little information on the transient behavior is available.

A transient analysis useful for practical circuit design is the objective of this investigation.

1.2 Requirements for Circuit Design

An optimum description of the transient response provides the circuit designer with:

- an accurate response time
- an accurate waveshape
- a solution for arbitrary termination (source and load)
- a solution for arbitrary drive (voltage and current)
- reasonable accuracy with minimum computational labor
- reasonable accuracy with minimum graphical construction
- a tool amenable to conventional circuit design
- a tool amenable to construction of charts for routine design
- a device model displaying the essential physical processes
- a device model supplying a qualitative feel for circuit behavior
- a set of approximations of increasing accuracy
- an estimate of error associated with each successive approximation.

The last two criteria deserve amplification. Circuit design frequently involves two steps: initially, an order of magnitude approach to assess feasibility of design, and subsequently, a more quantitive approach to make certain that the circuit meets specifications. Correspondingly, many design problems require models of two levels of accuracy: initially, a simple model for order-of-magnitude design, and subsequently a more adequate model for numerical evaluation. It is fortuitous that the simple model often gives results of adequate accuracy thus eliminating the need for subsequent refinement.

1.3 Background

Considerable analytic study has been given to distributed-parameter networks. To assess the applicability of this previous work, certain aspects of the problem under consideration are noted:

a) Distributed parameter networks with resistive and capactive elements only are of interest since in practice inductive effects are negligibly small. Thus, the wealth of material on transmission lines[5] containing an inductive component is not directly applicable.

b) Only the terminal response is needed. Much of transmission-line analysis, concerned with the more general problem of response as a function of distance along a line, provides unneeded information. On the other hand, the information it does provide is insufficient for our purposes.

c) For digital circuit design, transient response is of concern. A majority of the investigation of distributed parameter networks has been for sinusoidal excitation.

d) Weber[6] examined the transient response of a R-C transmission line, but the terminations considered are not of particular interest in circuit design.

e) The steady state response of the thin-film network (Fig.1) under consideration has been investigated[1,4] but no corresponding work on the transient response has been published.

As an alternative approach one may look for analogies to networks characterized by distributed resistance and capacitance. Since this type of network is governed by the diffusion equation, a large class of systems[7] governed by this equation are analogs. In particular, the flow of charge in a distributed parameter network is analogous to the flow of heat energy in a one-dimensional rod.

Two advantages result from exploiting this analogy: First, a conceptual clarification of the underlying physical phenomena is obtained. From a more practical point of view, the methods for solution of this classical problem,[8] including the powerful Laplace transform approach, are directly applicable to the distributed parameter network under consideration.

1.4 Proposed Approach

As a first step, the heat analogy is established and exploited to determine which method for solution of the heat problem is most amenable to the requirements of circuit design. It is found that a method using the Laplace transform yields a solution for arbitrary termination and drive, but only after considerable computational effort. We are thus motivated to search for a different approach.

Linvill[9] noted that concentration on the differential behavior of a device (or, what is equivalent, analysis of the distributed model of a device) leads not only to prohibitive computational labor, but also to results that are incompatible with the demands of flexible circuit design. Motivated by Linvill's pioneering work on lumped models for diodes and transistors, a lumped model of the distributed-parameter network is constructed.

In a companion paper[10] the versatility and power of the lumped model is demonstrated. The model is used to develop a design chart for use in repetitive design work and to predict the transient response of the microsystem in typical applications. The responses are verified experimentally showing the accuracy to be expected.

2. DIFFUSION ANALOGY

2.1 Analogous Systems

A comparison of the two analogous systems is summarized in Tables 1 to 6. The thermal system consists of a rod insulated on its lateral surfaces, thus constraining the flow of heat energy to one dimension. Analogous quantities are defined (Tables 1 to 3) to obtain generalized equations (Tables 4 to 6) for boundary conditions and physical laws. To clarify the terminology used

Table 1. Units

Unit	Thermal	Electrical
L	Meter	Meter
T	Second	Second
Q	Kilocalorie	Coulomb
V	°K or °C	Volt

Table 2. Definition of Analogous Physical Constants

Physical Constants	Thermal System	Electrical System	Units			
			L	T	Q	V
L	Length of Rod	Length of Film	1	0	0	0
A	Cross Sectional Area of Rod	Cross Sectional Area of Film	2	0	0	0
r	Resistance per Unit Length	Resistance per Unit Length	-1	1	-1	1
c	Capacitance per Unit Length	Capacitane per Unit Length	-1	0	1	-1
R = Lr	Resistance	Resistance	0	1	-1	1
C = Lc	Capacitance	Capacitance	0	0	1	-1
τ = RC	Time Constant	Time Constant	0	1	0	0

Table 3. Definition of Analogous Variables

Variables	Thermal System	Electrical System	Units			
			l	T	Q	V
$V(x,t)$	Temperature	Voltage	0	0	0	1
$i_{diff}(x,t)$	Heat Flux Density	Diffusion Current Density	-2	-1	1	0
$I_{diff}(x,t) = Ai_{diff}(x,t)$	Heat Flux	Diffusion Current	0	-1	1	0
$Q(x,t)$	Heat Energy Stored Per Unit Length	Charge Stored Per Unit Length	-1	0	1	0
$i(x,t)dx = \frac{\partial Q(x,t)}{\partial t}dx$ $= cdx \frac{\partial V(x,t)}{\partial t}$	Flow of Heat Energy Into Storage	Flow of Charge Into Storage	0	-1	1	0

Table 4. Typical Boundary Conditions at Input

Boundary Condition	Thermal System	Electrical System	Comments
$I_{diff}(0,t) = -\sigma A \frac{\partial v(0,t)}{\partial x} = u(t)$ where $u(t) = 1$ for $t > 0$ and $= 0$ for $t < 0$	Unit step of heat flux applied at the input	Unit step of current applied at the input	Easily handled, in principle, by conventional methods discussed by Crank.[7]
$V(0,t) = u(t)$	Unit step of temperature applied to input	Unit step at voltage applied to input	
$I_{diff}(0,t) = f(t)$ where $f(t)$ is an arbitrary function	Arbitrary function of heat flux applied at input	Arbitrary function of current applied at input	This could result from driving with a source of finite impedance. Such problems yield in principle to the method of Gray (Sec. 2.3).
$V(0,t) = f(t)$	Arbitrary function of temperature applied at input	Arbitrary function of voltage applied at input	

Table 5. Boundary Conditions at Output

Boundary Condition	Thermal System	Electrical System	Comments
$I_{diff}(L,t) = - \sigma A \frac{\partial V(L,t)}{\partial x} = 0$	End of rod insulated	Open circuited output	
$V(L,t) = 0$	Temperature at end of rod held at zero	Short circuited output	
$I_{diff}(L,t) = f(t)$	Arbitrary heat flux at the output	Arbitrary current flow at output	This could result from termination with arbitrary load impedance. Such problems yield, in principle, to the method of Gray (Sec. 2.3).
$V(L,t) = f(t)$	Arbitrary temperature function at the output	Arbitrary voltage function at the output	

Table 6. Physical Laws of Differential System

Physical Law	Thermal System	Electrical System	Equation
Diffusion flow	Flow of heat energy due to diffusion is proportional to the temptemperature gradient at point x and time t.	Flow of electric charge due to diffusion is proportional to voltage gradient at point x and time t.	$I_{diff}(x,t) = -\sigma A\frac{\partial V(x,t)}{\partial x}$
Storage	Heat energy stored in differential volume about point x at time t.	Electric charge stored in differential volume about point x at time t.	$Q(x,t)dx = cdx\ V(x,t)$
Storage flow	Time rate increase of heat energy stored in differential volume about point x at time t. Flow of heat energy to storage.	Time rate increase of electric charge stored in differential volume about point x at time t. Flow of current into storage.	$\frac{\partial Q(x,t)}{\partial t}dx = cdx\ \frac{\partial V(x,t)}{\partial t}$
Continuity	Net flow of heat energy into differential volume about point x at time t equals the time rate increase of heat energy stored in the same differential volume.	Net flow of electric charge into differential volume about point x at time t equals the time rate increase of charge stored in the same differential volume.	$\frac{\partial I_{diff}(x,t)}{\partial x} = -\frac{c\partial V(x,t)}{\partial t}$

in these tables, it may be noted that the terms "convection" or "drift" are usually associated with the flow of electric charge arising from a voltage gradient. However, because of the analogy being drawn between temperature and voltage, it appears that "diffusion" is the most appropriate term to use in the present investigation.

The thermal and electrical systems have common mathematical formulation because phenomena in both systems are special cases of a more general physical process, the diffusion process, in which flow is proportional to the concentration gradient. The quantity which flows may be mass, charge, energy, or any other physical quantity subject to the laws of conservation and continuity. Thus, the same differential equations also applies for such diverse problems as molecular diffusion in physical chemistry and the flow of minority carriers across a diffusion transistor base in which recombination is negligible.

2.2 Exploiting the Analogy

To facilitate the conversion from one system to another, Tables 1 through 6 can be extended to make available the simultaneous solution of a wide variety of diffusion problems. Such a scheme is useful conceptually because some phenomena are more evident in one system than another. For example, the fact that there is a time delay in a gas diffusion problem is perfectly clear; it is perhaps a somewhat less obvious characteristic of the distributed parameter network.

However, the principal value of establishing the analogy is that it provides the circuit analyst with the highly developed mathematical approach to diffusion; both techniques and solutions become available. A survey of the literature of the Laplace transform, is therefore advisable.

A systematic and effective method for obtaining the transient response of a distributed-parameter network terminated in arbitrary source and load impedances is suggested by Gray[11] who derived sets of calculated curves for the transient response of a heat pump. Although this problem is not an exact analog to the one under consideration, the similarity of the mathematical formulation involved warrants use of the Gray method as a guide.

2.3 The Laplace Transform Method

Although the behavior of a heat pump is not exactly analogous to that of the thin-film structure, the approach proposed by Gray gives a powerful procedure for finding the transient response of the thin-film system with arbitrary source and load termination. In fact, this procedure was used by Starner,[12] in a preliminary investigation of the transient response of thin-film structures. The procedure is:

1) Solve the diffusion equation by assuming exponential variation with time for I and V.

2) Express these solutions in terms of the terminal variables. Holm[1] has shown that for the distributed-parameter circuit (Fig. 1), steps 1 and 2 lead to the matrix equation:

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \frac{R}{\theta} \begin{bmatrix} 1/\tanh \theta & 1/\sinh \theta \\ 1/\sinh \theta & 1/\tanh \theta \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}$$

where $\theta = (i\omega RC)^{1/2}$. It is useful to generalize Holm's definition of θ by substituting $s = \sigma + i\omega$ for $i\omega$. The justification for this follows simply from assuming e^{st} rather than $e^{i\omega t}$ variations.

3) Using two-port theory, write the desired transfer function as combinations of these transcendental functions and the source and load impedances.

4) The output transform is determined by multiplying the transfer function by the input transform. It is now necessary to transform to the time domain.

5) To obtain the inverse transform, find the poles of the output transform by numerical or graphical evaluation. These poles are the roots of a transcendental equation; hence there are an infinite number of them.

6) Compute the residue at each pole by application of L'Hospital's rule to the output transform.

7) By Cauchy's residue theorem, the output function of time is the sum of the residues.

Since there is an infinite number of poles associated with the output transform, the resulting time function is in the form of an infinite series. To yield results of any use such a series must be truncated and the response graphically constructed.

Gray[11] presents an example of this method showing the manipulations and the difficulties encountered.

2.4 Limitations of Laplace Transform Method

Although the inverse transform method, in principle, gives a solution for arbitrary termination and drive, it does not meet the needs for flexible circuit design. The time required to obtain a solution may be reduced by the use of computers, but considerable computational effort is still necessary. Furthermore, the resulting solution does not lend itself to physical interpretation nor does it supply the insight so necessary to the practical circuit designer.

A comparison of the results of this method with the optimum description of transient response for design purposes (Section 1.2) motivates the search for a new approach.

3. MODELS

3.1 A Distributed Model

The technique of evaluating the transient response by use of the inverse

Laplace transform is based on a distributed model of the structure. Such a model, emphasizing differential behavior, leads to partial differential equations (Table 6) and corresponding transcendental solutions (Sec. 2.3). The essential features of this model are

a) At each point two types of current occur:

diffusion current across a finite area A

$$I_{diff}(x,t) = -\sigma A \frac{\partial V}{\partial x}(x,t)$$

Storage current or time rate increase of charge stored in differential volume Adx:

$$i(x,t)\,dx = \frac{\partial Q(x,t)}{\partial t}\,dx = c\,dx\,\frac{\partial V(x,t)}{\partial t}$$

b) At each point, a continuity relation holds: Net current into the differential volume equals the time rate increase of charge stored, that is, the storage current. Hence

c) At each point, the diffusion equation, derived from relations above, applies:

$$\frac{\partial^2 V(x,t)}{\partial x^2} = \frac{\tau}{L^2}\frac{\partial V(x,t)}{\partial t}$$

The distributed model is an idealization of the actual structure. It neglects:

- contact resistance
- stray capacitances and inductances
- inhomogeneity in the materials
- finite resistance of the conductive film
- surface phenomena

3.2 The Purpose of Models

The model of a device or a process is an idealization that lends itself to mathematical analysis. In the construction of a model one generally emphasizes first-order and neglects second-order effects. For example, to make the problem of heat flow in a rod mathematically tractable, one demands that the model of the process assume

- the material is homogeneous
- the temperature is the same at any given cross section
- the thermal conductivity is not a function of temperature
- radiation may be neglected

With the development of a model, attention shifts from actual device behavior to model behavior. The correspondence between actual performance and model performance depends, of course, on the approximations made in deriving the model.

In each engineering application, specific requirements determine the desired degree of correspondence. An example of the seemingly gross idealizations design engineers make is illustrated by the first-order model, often chosen for a common-base transistor amplifier. Such a model (Fig. 2) emphasizes the essential features--low input, high output impedance and almost unity current transfer--without giving detail which frequently is unnecessary and undesirable in the initial stages of design.

4. LUMPED MODELS

4.1 Properties

Since the distributed model and its associated inverse transform method is unsatisfactory for circuit design, a change of model is suggested. In the new model (Fig. 3) voltage, instead of being a continuous variable, is defined at only three points--the two terminals and the middle of the model.

Comparison of the essential features of the two models given in Table 7 shows that:

a) Partial differential equations are exchanged for ordinary differential equations.

b) Since only diffusion current is allowed at the end nodes, all charge storage occurs at the middle node.

c) Both models store the same charge in the steady state as the physical structure and have an equivalent parallel plate capacitance of $\epsilon WL/a$.

d) Both models have the same steady-state resistance as the physical structure namely $L/(\sigma\ kW)$.

e) Both models, as well as the actual structure, are symmetrical with respect to input and output.

Here W denotes film width, k denotes film thickness, L gives the film length, σ is the conductivity of the film and ϵ is the dielectric constant of the wafer.

4.2 Equivalent Circuit

An electrical circuit equivalent to the lumped model is given in Fig. 4. Future reference to the lumped model will apply to its equivalent circuit, since a simple electric network is more easily dealt with than a set of laws. Note that an analogous equivalent circuit for the distributed model requires an infinite number of elements.

This equivalent circuit could have been derived by a different approach. The distributed model corresponds to two-port parameters described by hyperbolic functions (Sec. 2.3), which may be approximated by finite power series (Table 8). The first terms of the series corresponds to the lumped model discussed above. Higher-order lumped models correspond to several terms of the power series. In Table 8, $\theta = (sRC)^{1/2}$ where s is the complex frequency variable.

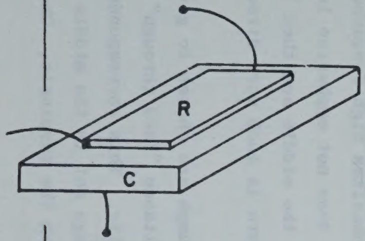
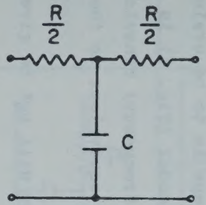
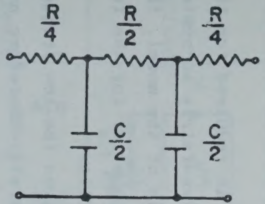
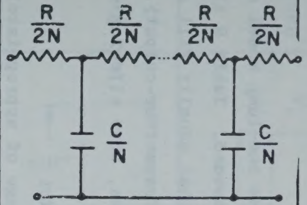
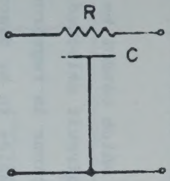
4.3 Higher-Order Models

The first-order approximations derived above (Sec. 4.1) and its equivalent circuit (Sec. 4.2) will be shown useful for a wide variety of circuit applications. The limitations of the simple model become noticeable if:

Table 7. Comparison Between Lumped and Distributed Models

Type of Model	Distributed	Lumped		
Where Applicable	At each point	At input node	At middle node	At output node
Diffusion Current: $I_{diff}(x,t)$	$-\sigma A\frac{\partial V(x,t)}{\partial x}$	$\frac{-\sigma A}{L/2}(V-V_{in})$	$\frac{-\sigma A}{L/2}(V-V_{in})$	$\frac{-\sigma A(V_{out}-V)}{L/2}$
Storage Current: $i(x,t)\ dx$	$c dx\ \frac{\partial V(x,t)}{\partial t}$	None	$\frac{\epsilon LW}{a}\ \frac{dV}{dt}$	None
Continuity Relations:	$\frac{-\partial I_{diff}(x,t)}{\partial x} = c\frac{\partial V(x,t)}{\partial t}$	$I_{source} = I_{in}$	$I_{out} - I_{in} = C\frac{dV}{dt}$	$I_{out} = I_{Load}$

Table 8. Power Series Justification of Lumped Model

				
Equivalent Circuit:	1 Lump	2 Lumps	N Lumps	Distributed
Successive Approximations	1st Approximation	2nd Approximation	Nth Approximation	Exact
$Ry_{11} = Ry_{22}$	$\frac{1 + \theta^2/2}{1 + \theta^2/4}$	$\frac{1 + \theta^2/2 + \theta^4/32}{1 + 3\theta^2/16 + \theta^4/128}$	$\frac{1 + \theta^2/2 + \theta^4/24}{1 + \theta^2/6 + \theta^4/120}$	$\frac{\cosh \theta}{\theta^{-1} \sinh \theta}$
$Ry_{12} = Ry_{21}$	$\frac{1}{1 + \theta^2/4}$	$\frac{1}{1 + 3\theta^2/16 + \theta^4/128}$	$\frac{1}{1 + \theta^2/6 + \theta^4/120}$	$\frac{1}{\theta^{-1} \sinh \theta}$

the driving function and circuit environment emphasize those aspects of the actual structure which are least accurately represented by the model (Example 2, Sec. 4.4).

the network has a frequency response which cannot be approximated in general shape by the type of lumped circuit selected.

accuracy in numerical design calculations is required which exceeds the inherent accuracy of the model. Examples to be discussed give accuracies of within 10 per cent.

For cases in which the limitations of the one-lump model are important, a judicious choice of a higher-order model is needed. Table 9 illustrates an example, the derivation of a two-lump model that usually will give sufficient accuracy for use in decoupling applications (capacitor-output). Note that this equivalent circuit has two natural frequencies, thus allowing description of a response which passes through a maximum.

4.4 Circuit Environment and Choice of Model

Two examples will illustrate the advantage of appropriately selecting the lumps of a model:

Example 1: The response to a step of voltage is required for the configuration of Fig. 5. The simple model is expected to give fairly accurate response time. Although all the charge is stored in the middle of the model, this will tend to distort only the short time (high frequency) behavior for the following reasons:

- the device is essentially a low-pass filter
- the Fourier spectrum of a step emphasizes the low frequencies.

The simple approximation to the storage effect will therefore not be important and use of the simple model of Fig. 5 is justified.

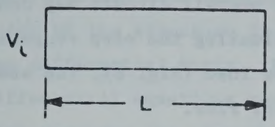
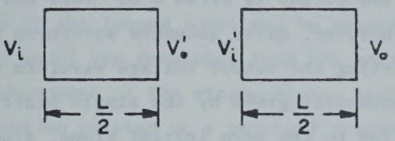
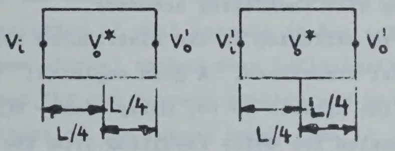
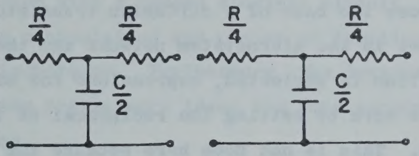
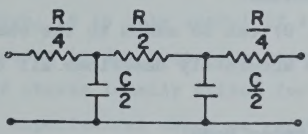
Example 2: The response to an impulse of voltage is to be evaluated for the configuration of Fig. 6. The simple one-lump model (Fig. 6) is useful in initial stages of design. The long time (low frequency) behavior will be faithfully portrayed by this model since it is equivalent to the quite accurate, distributed model for dc excitation.

On the other hand, the short time response will not be given accurately for the following reasons:

- The configuration of the device emphasizes high frequencies.
- The Fourier spectrum of an impulse does not emphasize low frequencies.
- In the model, current is passed to the middle and then "diffuses" to the ends; in the actual device, there is much more direct "feed through" to the load.

To improve the short time behavior, lumps are chosen to give the following model (Fig. 7), which emphasizes the capacitance "feed through" to the ends and still maintains a good deal of simplicity. This corresponds to choosing lumps as in Table 9, but defining the voltage not in the middle of each lump, but rather one-quarter of a lump length from the terminals. Choice of the

Table 9. Construction of a Two-Lump Model

	Representation	Operation
Actual Structure		Define terminal variables
Two-lump models defined	 V_i V_o' V_i' V_o $\frac{L}{2}$ $\frac{L}{2}$	Break structure into two lumps. Define terminal variables of each lump
Variables of two-lump model defined	 V_i V_o^* V_o' V_i' V_o^* V_o' $L/4$ $L/4$ $L/4$ $L/4$	Define all pertinent variables treating each lump separately
Equivalent Circuit of two lumps	 $\frac{R}{4}$ $\frac{R}{4}$ $\frac{C}{2}$ $\frac{R}{4}$ $\frac{R}{4}$ $\frac{C}{2}$	Draw equivalent circuit for each lump.
Two-lump circuit	 $\frac{R}{4}$ $\frac{R}{2}$ $\frac{R}{4}$ $\frac{C}{2}$ $\frac{C}{2}$	Cascade networks to conserve continuity.

model is somewhat arbitrary. In practice, a good choice strongly depends on how familiar the designer is with the nature and requirements of the specific problem.

5. DISCUSSION

5.1 Effect on Waveshape

The simple model has one storage element and hence one natural frequency associated with it, although the overall circuit may contribute natural frequencies as well. Hence, in evaluating the step response for the common capacitor configuration with resistance load (Fig. 8), the one-lump (or one-pole) model will give simple exponential rise.

Since it is evident that the voltage must diffuse across the actual device, an initial condition at the output is $dv/dt = 0$. This fact is not disclosed by the model. Frequently, however, quite accurate waveforms can be sketched on a linear time scale by starting the output voltage waveform with zero slope and then superposing the exponential given by the simple model. Higher order models give a better approximation to the zero initial slope, since the second and higher order natural frequencies will tend to match the slope as well as the amplitude of the response with increasing accuracy.

The delay time due to diffusion in the distributed structure was found experimentally for several structures. A good empirical "rule of thumb" for the magnitude of delay time appears to be: delay time ~ RC/10.

An analytical expression for delay resulting from the finite time it takes carriers to travel across the base of a diffusion transistor was derived by Chow.[13] Since processes in the microsystem network and the transistor base are analogous if recombination is neglected, expressions for microsystem delay can be obtained from Chow's work by setting the reciprocal of the minority-carrier lifetime equal to zero. This is not done here because the delay will not be significant in most design problems.

An ideal delay line (Fig. 9) can be added to the one lump model to give a circuit which frequently quite accurately describes all the essential features of the output waveform.

5.2 Justification of the Lumped Model

It is instructive to compare the alternative approaches in a block diagram (Fig. 10) noting.

1) when approximations are made
2) how the need for higher approximations is determined
3) the relative susceptibility to computational error.

The essential difference between the lumped model and inverse transform approach is in the "timing" of approximations. Starting with the actual structure, both methods neglect secondary effects and arrive at a set of laws describing the phenomena of prime importance. Here the attacks diverge. The inverse transform method aims at the solution of the partial differential equations to yield the desired output transform in terms of combinations of transcendental functions. A numerical or graphical evaluation of the poles of the output transform allows use of the Cauchy residue theorem to give a solution in the form of an infinite series. The series is truncated and the solution constructed graphically. Note that an approximation is made in the last step of analysis in order to get a usable result.

In the lumped-model approach, approximation is made much earlier to avoid working with partial differential equations and their transcendental solutions. The distributed character of the structure is ignored by treating the phenomena on a finite rather than differential basis. This leads to ordinary rather than partial differential equations and to a simple, and hence practical, equivalent circuit which provides the circuit designer with all the tools of lumped network theory.

The justification of the lumped model may be summarized as follows:

(a) The lumped model was developed from the morphology, physical constants, and basic physical phenomena of the microsystems; each part of the corresponding equivalent circuit can be related to a physical process. In the one-lump model, current "diffuses" from the input terminal to the middle node where part of the charge is stored and the remainder continues its "diffusion" to the output terminal. This is a description of what happens in the actual device, but on a finite (and hence approximate) rather than infinitesimal basis. Such a conceptual scheme bridges the gap between physics and model.

(b) The usefulness of the lumped model lies in its replacement of an infinite-element system with a lumped electric circuit. Thus, the problem of design is reduced to manipulation and choice of familiar components, utilizing standard synthesis techniques. Furthermore, the designer has available both a simple model to form preliminary ideas and more accurate models to check design specifications.

(c) Although the model provides the insight and tools necessary for effective circuit design, it is only useful if it gives reasonably accurate results. It will be shown in the companion paper that the model is amenable to the construction of charts ideally suited for routine design work; furthermore, a comprehensive experimental check on response time and wave shape resulting from various orientations and terminations is made to ascertain that the accuracy obtained meets the demands of the circuit designer. Thus, the model satisfies the criteria for description of the transient response (Sec. 1.2) and justification is complete.

A tutorial review of models[14] useful for determining the transient response of transistors

a) derives, independently, several alternative models

b) shows in what respects equivalence among models exists.

c) compares the models as tools for analysis and physical understanding.

clarifying the principles involved in establishing practical models for determining the transient response, it parallels the objective of this investigation.

REFERENCES

1. C. Holm, "Distributed Resistor-Capacitor Networks for Microminiaturization," R.R.E. Technical Note No. 654, Worchester, England, March 1960.
2. P. S. Castro and W. W. Happ, "Distributed Parameter Circuits and Microsystem Electronics," Proc. Nat'l. Electronics Conference, Vol. 16, p. 448-460, 1960.
3. T. R. Nisbet and W. W. Happ, "Jacobians of Frequency-Selective Networks," Electronic Industries, Vol. 20, p. 86-94, January 1960.
4. P. S. Castro and W. W. Happ, "Distributed Parameter Circuit Design Techniques," Proc. Nat'l Electronics Conf., Vol. 17, Oct. 1961.
5. L. Brillouin, "Wave Propagation in Periodic Structures," McGraw-Hill Book Co., 1946.
6. E. Weber, "Linear Transient Analysis," Vol. 2, John Wiley and Sons, Inc., 1956.
7. J. Crank, "The Mathematics of Diffusion," Oxford, Clarendon Press, 1956.
8. H. S. Carslaw and J. C. Jaeger, "Conduction of Heat in Solids," Oxford, Clarendon Press, 2nd edition, 1958.
9. J. Linvill, "Lumped Models of Transistors and Diodes," Proc. I.R.E., Vol. 46, pp. 1141-1152, June 1958.
10. W. W. Happ and F. A. Lindholm, "Design Charts for the Transient Response of R-C Distributed Parameter Networks," (companion paper to follow in these Proceedings), 1963.
11. P. E. Gray, "The Dynamic Behavior of Thermoelectric Devices," Technology Press of MIT and John Wiley and Sons, Inc., 1960.
12. D. L. Starner, "Transient Response of Some Distributed Parameter Configurations, Part 1," LMSD 703095 (internal communication), Lockheed Missiles and Space Division, Palo Alto, Calif., Sept. 1960.
13. W. F. Chow and J. J. Suran, "Transient Analysis of Function Transistor Amplifiers," Proc. IRE, Vol. 41, pp. 1125-1129, Sept. 1953.
14. D. J. Hamilton, F. A. Lindholm, and J. A. Narud, "Large Signal Nodels for Junction Transistors," Digest of Technical Papers, p. 55-56, International Solid-State Circuits Conference, Philadelphia, Pa., 1963.

LIST OF FIGURE CAPTIONS

Figure 1. The Thin-Film Structure

(a) Schematic of Actual Structure

(b) Network Symbol

Figure 2. (a) Common-base Amplifier

(b) First-order Model

Figure 3. Definitions of Variables for One-Lump Model

Figure 4. One-Lump Equivalent Circuit

Figures 5 through 9.

No Captions

Figure 10. Comparison of Approaches

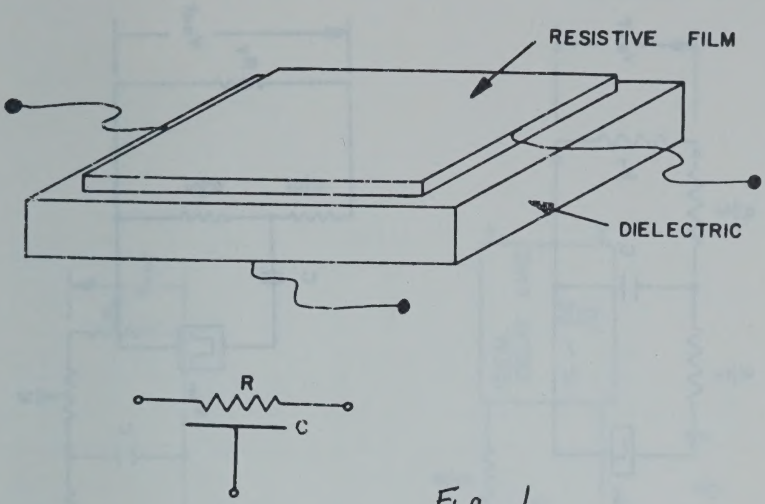

RESISTIVE FILM
DIELECTRIC
R
C

Fig. 1

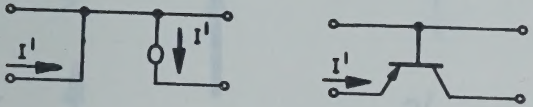


Fig. 2

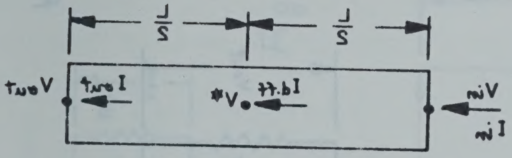


Fig. 3

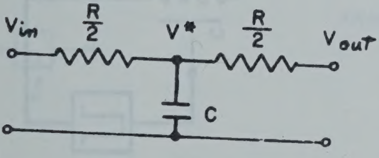

Vin
R/2
V*
R/2
Vout
C

Fig. 4

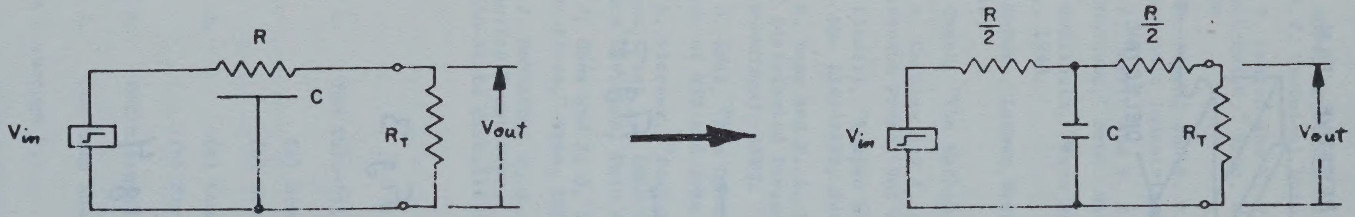

R
C
V_in
R_T
V_out
R/2
R/2
V_in
C
R_T
V_out

Fig. 5

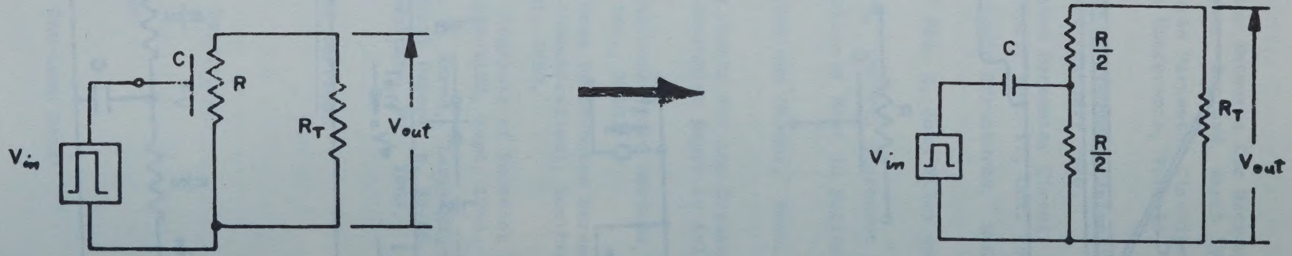

C
R
V_in
R_T
V_out
C
R/2
R/2
V_in
R_T
V_out

Fig. 6

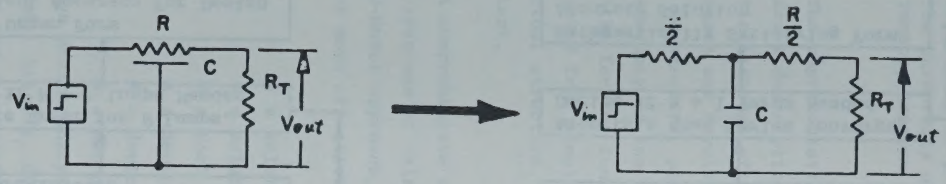

R
C
R_T
V_in
V_out
R/2
R/2
C
R_T
V_in
V_out

Fig 8

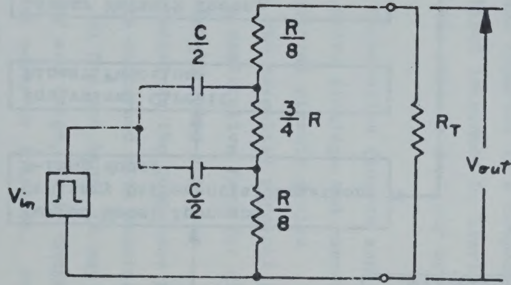

C/2
R/8
3/4 R
R_T
V_out
V_in
C/2
R/8

Fig. 7

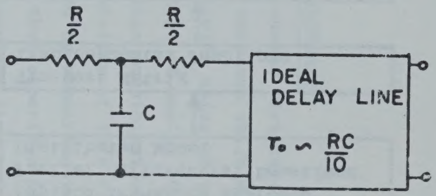

R/2
R/2
C
IDEAL DELAY LINE
T₀ ∼ RC/10

Fig. 9

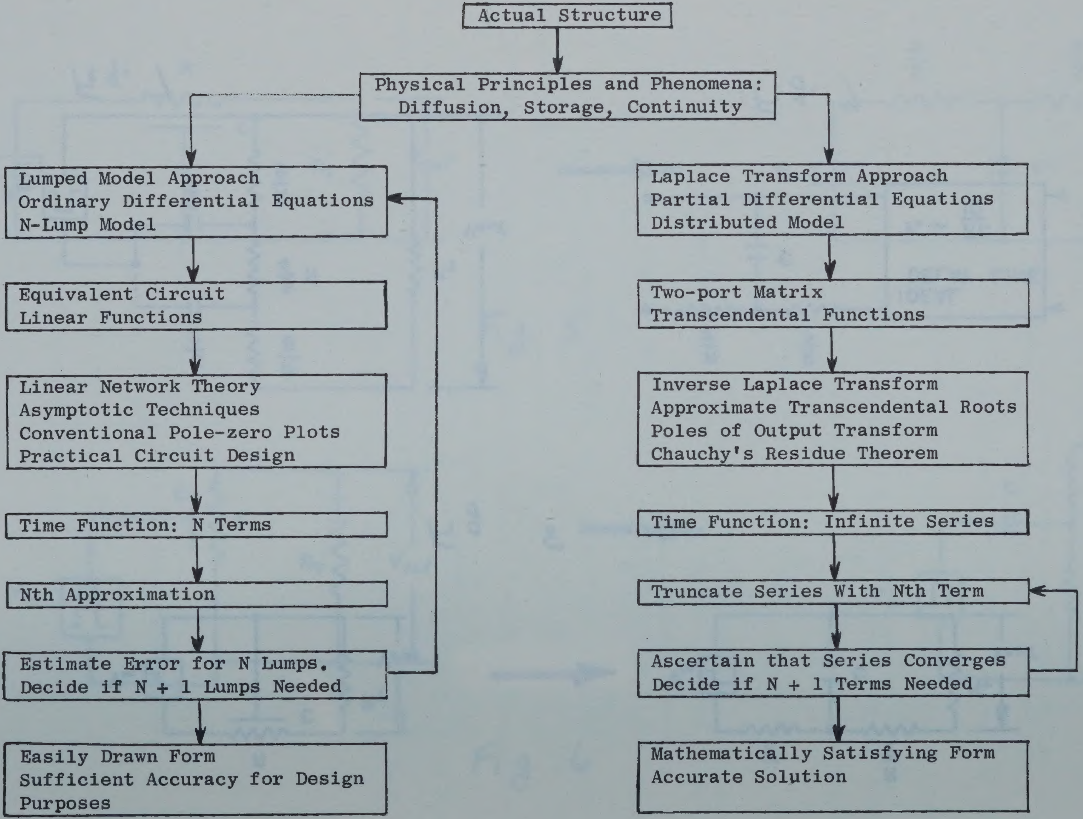

Actual Structure
Physical Principles and Phenomena: Diffusion, Storage, Continuity
Lumped Model Approach
Ordinary Differential Equations
N-Lump Model
Equivalent Circuit
Linear Functions
Linear Network Theory
Asymptotic Techniques
Conventional Pole-zero Plots
Practical Circuit Design
Time Function: N Terms
Nth Approximation
Estimate Error for N Lumps.
Decide if N + 1 Lumps Needed
Easily Drawn Form
Sufficient Accuracy for Design Purposes
Laplace Transform Approach
Partial Differential Equations
Distributed Model
Two-port Matrix
Transcendental Functions
Inverse Laplace Transform
Approximate Transcendental Roots
Poles of Output Transform
Chauchy's Residue Theorem
Time Function: Infinite Series
Truncate Series With Nth Term
Ascertain that Series Converges
Decide if N + 1 Terms Needed
Mathematically Satisfying Form
Accurate Solution

Fig. 10.

DESIGN CHARTS FOR THE TRANSIENT RESPONSE OF R-C DISTRIBUTED PARAMETER NETWORKS

W. W. HAPP
College of Engineering Sciences
Arizona State University
Tempe, Arizona

and

F. A. LINDHOLM
Department of Electrical Engineering
The University of Arizona
Tucson, Arizona

ABSTRACT

The representation of distributed-parameter structures by lumped models provides a description of their transient response which makes feasible analytical design of pulse and digital circuits using thin-film components. Techniques are developed for the construction of design charts which, for given constraints, pictorially display design possibilities.

Adequacy of the lumped-model representation of thin-film structures for step-like forcing functions is investigated by comparing the transient response calculated from simple lumped models with that experimentally obtained from the actual structure. Agreement in both response time and waveshape is excellent.

INTRODUCTION

A comprehensive study of alternative approaches for determining the transient response of thin-film structures was reported in a companion paper.[1] A lumped-model approach, similar to that first introduce by Linvill,[2] was shown to meet most effectively the demands of the circuit designer by providing him with:

- a solution for arbitrary termination (source and load)
- a solution for arbitrary drive (voltage and current)
- a tool amenable to conventional circuit design
- a device model displaying the essential physical processes
- a device model supplying a qualitative feel for circuit behavior
- a set of approximations of increasing accuracy

The device analyst will find the material in the companion paper of main interest. Briefly, the companion paper is concerned with examining a heat analogy and constructing lumped models. On the one hand, it may be regarded as somewhat of a tutorial review since, by and large, it is comprised of material which is available in either text books or journals. On the other hand, a rather general approach to the problem of model development is described which may be applicable for new problems of a similar nature. Thus, the companion paper may contain information of more lasting, although less immediate, importance.

The purpose of this investigation is to show that a simple lumped-model representation of thin-film structure will provide:

a) a tool amenable to the construction of charts for routine design

b) a reasonable approximation to actual response time

c) information on waveshape of importance to the designer

The study presented here is self-contained and is intended for the practical circuit designer primarily interested in using the lumped model as a design tool.

1. CONSTRUCTION OF DESIGN CHARTS

1.1 Statement of the Problem

The number of different boundary conditions, and initial conditions that may be applicable in different circuit configurations is overwhelming and no attempt will be made to catalogue them. However, circuit design can often be facilitated by construction of a chart which gives at a glance useful design parameters, thus replacing repetitive computations.

As an example, a design chart will be developed for the application shown in Figure 1, in which the step voltage response of the structure terminated in a series combination of resistance and capacitance must be evaluated. The design chart will serve as a guide for initial design, and considering the tolerance of circuit components commercially available, may also suffice as a final design tool in many cases.

The thin-film structure denoted by its network symbol in Figure 1 consists of a rectangular dielectric wafer upon which a resistive film has been deposited on one side and conducting film on the other. This structure has been extensively analyzed[3,4,5,6,7] for steady-state applications. Representative values of dimensions and physical constants are given in Table 1.

Table 1. Representative Values for a Thin-film Structure

Symbol	Quantity	Representative Value
L	length of film	10^{-2} m
w	width of film	10^{-3} m
k	thickness of film	10^{-7} m
a	thickness of wafer	10^{-3} m
ϵ	dielectric constant of wafer	10^{-8} farad/m
σ	conductivity of film	10^{-3} mhos -□

A "zero-order" approximation to the time response of the unloaded structure is obtained from the dimensions and constants by use of familiar relations:

$$R = \frac{1}{\sigma}\,\frac{L}{kw} \text{ and } C = \epsilon\,\frac{Lw}{a}$$

and therefore,

$$RC = \epsilon L^2/(\sigma\ ka).$$

This RC time constant will provide a convenient normalizing factor for the response time computed from the lumped models.

1.2 Topological Approach

Figure 2 shows the one-lump model terminated by resistance and capacitance. Since there are two capacitors in the entire circuit, there will be two natural frequency components. As a first step in developing a desing chart for any termination of this type, an asymptotic description of the two natural modes will be found. The notation to be used is

$RC = \tau$ = distributed-parameter network time constant

τ_L = time constant of the low frequency mode

L = amplitude of low frequency mode

τ_S = time constant of high frequency mode of entire circuit

S = amplitude of high frequency mode

τ_c = time required for output voltages to rise to $(1-\epsilon^{-1})$ or about 63 per cent of its final value

R_T = terminating resistance

C_T = terminating capacitance

Consider the following terminations:

a) series combination of R_T and C_T

b) pure capacitance C_T, a special case of (a) for $R_T = 0$.

c) pure resistance R_T; this cannot be treated as a special case of (a), and is discussed in Sec. 1.3.

The response time τ_c results from the composite behavior of two interacting circuits, in particular from τ_L and τ_S and L and S, defined above. However, one of the two time constants dominates for a number of asymptotic cases, namely:

R_T much greater or much smaller than R

C_T much greater or much smaller than C

R_TC_T much greater or much smaller than RC

For these asymptotic cases, the resistance and capacitance of a single loop essentially determines the time response. This loop is found from the topology of the network by including only those elements which cannot be replaced in effect by short or open circuits. Application of this topological argument to the lumped circuit under consideration gives the results shown in Table 2 for a voltage step and summarized in Table 3. For a current step the response time is given by

$$\tau_c = \frac{C_TC}{C+C_T}\left(\frac{R}{2} + R_T\right)$$

and the information in Table 4 corresponds to that given in Table 3.

With considerable practice, the topological approach described above gives an estimate of asymptotic circuit behavior at a glance; however, it is essential to strengthen analytically the estimates obtained by topological reasoning.

Table 3. Voltage Step

C_T/C R_T/R	large	small	absent
large	$\tau_c/\tau = 1/2$	1/2	1/2
small	C_T/C	1/2	1/4
short	C_T/C	1/2	1/4

Table 4. Current Step

C_T/C R_T/R	large	small	short
large	$\tau_c/\tau = R_T/R$	$(R_TC_T)/(RC)$	R_T/R
small	1/2	$C_T/(2C)$	1/2
short	1/2	1/2	1/2

1.3 An Apparent Anomaly: Resistive Load

Although $C_T = 0$ is equivalent to an open circuit, the usual concept that infinite capacitance corresponds to a short circuit does not apply for calculation of the response time, because such a concept demands mathematically infinite response time for a purely resistive load which is not justified physically.

A simple argument explains the apparent anomaly. Starting with a finite parallel plate capacitance, for convenience, the capacitance may be reduced toward zero or made to approach infinity by two corresponding operations. By pulling the plates farther and farther apart, the capcitor approaches the usual physical notion of an open circuit. However, pushing the plates closer and closer together gives a structure which is capable of charge storage, *if* the two plates do not touch. Since we may only approach infinity, never quite getting there, the almost-infinite capacitor has one property that a short circuit does not have, charge storage. Hence, although the correspondence of an almost-infinite capacitor and a short circuit is valid for steady state behavior since neither may have a voltage developed across them, such a concept is misleading for evaluation of transient behavior since it takes a very long time for charge to be collected in such a structure.

From the equivalent circuit, the response time for a purely resistive load is

$$\tau_c = \frac{(R/2 + R_T)RC}{2(R + R_T}$$

1.4 Response Time for Arbitrary Load

Formulas for the response time of an arbitrarily loaded circuit will be useful to determine:

(a) asymptotic behavior and resultant design trends

(b) regions of validity and limitations of asymptotes

(c) critical points outside the region of validity of the asymptotes

To obtain τ_c as a function of R and C, determine the characteristics frequencies by setting the determinant of the impedance matrix of the circuit of Fig. 3 equal to zero:

$$\begin{vmatrix} R/2 + (pC)^{-1} & -(pC)^{-1} \\ -(pC)^{-1} & R/2 + R_T + (pC)^{-1} + (pC_T)^{-1} \end{vmatrix} = 0$$

Thus

$$p_{1,2} = - \frac{\frac{C_T}{C}\left(\frac{R_T}{R} + 1\right) + \frac{1}{2}}{\frac{C_T}{C}\left(\frac{R_T}{R} + \frac{1}{2}\right)} \left[1 \pm \sqrt{1 - \Delta}\right] \frac{1}{RC}$$

where

$$\Delta = 2 \; \frac{C_T}{C} \left(\frac{R_T}{R} + \frac{1}{2}\right) \left[\frac{C_T}{C} \left(\frac{R_T}{R} + 1\right) + \frac{1}{2}\right]^{-2}$$

and Δ has a maximum of

$$\Delta(\max) = .67 \text{ for } R_T \approx R \text{ and } C_T = C/2$$

Thus it is permissible to set

$\sqrt{1-\Delta} = 1 - \Delta/2$ with less than 15 per cent error even for $\Delta(\max)$ and with negligible error for all asymptotic cases. This gives two roots

$$p_1 = - 1/\tau_L \quad \text{and} \quad p_2 = - 1/\tau_S$$

which are the negatives of the reciprocal time constants

$$\tau_L = \left[(C_T/C)(R_T/R + 1) + 1/2\right] \tau$$

$$\tau_S = \left[(C_T/2C)(R_T/R + 1/2) / \tau_L\right] \tau^2$$

such that $\tau_S/\tau_L = \Delta/4$ is always less than 17 per cent.

Asymptotic limits of τ_L and τ_S give three distinct values for large or small values of R_T/R and C_T/C as plotted in Figure 4. The three regions are bounded by zones in which the asymptotic formulas do not necessarily hold.

1.5 Evaluation of Amplitudes

For the circuit of Figure 1, the response to a voltage step input is

$$v_o(t) = 1 + L \exp(-t/\tau_L) + S \exp(-t/\tau_S)$$

since $v_o(\infty) = 1$.

Further

$$v(0) = 0 \text{ yields } L + S = 1 \text{ and}$$

$$\frac{dv(0)}{dt} = \frac{4R_T}{RC(2R_T + R} = \frac{L}{\tau_L} + \frac{S}{\tau_S}$$

Calculation of L/S as a function of R_T/R and C_T/C gives the two asymptotic regions, shown in Figure 5, namely:

$$L > 10\ S \text{ and } S > 10\ L$$

as well as the boundary region for intermediate values.

1.6 Asymptotes for Design Charts

The time constants τ_L and τ_S and the corresponding amplitudes L and S plotted in Figure 4 and Figure 5 determine the circuit response τ_c shown in Figure 6. Clearly, if L is large, τ_c will approach τ_L conversely if S is large, τ_c will approach τ_S; therefore, two regions result in Figure 6 which are separated by a boundary zone and which correspond to values given in Tables 2 and 3 for asymptotic approximations. The resulting asymptotes form the skeleton for construction of the design chart, Figure 7. Note that, for a wide range of design applications, the information provided by a design chart consisting of asymptotes only is satisfactory.

1.7 Critical Points for Design Charts

It is advantageous to identify regions in the design charts for which asymptotic approximations do not hold within the limits of the desired or specified accuracy. For the model here considered, values not obtainable from asymptotic approximation are in the regions $R \sim R_T$ and $C \sim C_T$ shown as shaded zones in Figure 6. Having established a range of critical points in the nonasymptotic regions of the design chart, one may determine a set of critical points for τ_c by numerical point-by-point evaluation as follows:

(a) compute τ_L and τ_S for a given termination
(b) compute L and S
(c) compute $v_o(t)$
(d) compute τ_c such that $v(\tau_c) = .63$

The design chart shown in Figure 7 consists of the asymptotes previously described and the loci in non-asymptotic regions which are suggested by the chosen set of critical points. For many practical design problems, very few critical points, strategically placed, will be satisfactory.

The illustrative example just presented indicates that a simple lumped-model representation of thin-film structures admits the development of techniques for the rapid construction of charts for use in the design of pulse and digital circuits. For specified constraints (e.g., circuit rise time) a chart of this type indicates design possibilities (e.g., permissible values of R_T and C_T), thereby providing the designer with a flexible and effective graphical design tool.

2. EXPERIMENTAL PROCEDURE AND RESULTS

2.1 Description of Device

Methods of fabrication[4] previously reported by Fuller lead to the dimensions and characteristics of the experimental device shown in Table 5.

Table 5. Experimental Device

R	5.7 k	Measured on Impedance Bridge
C	4200 μμf	Measured on Impedance Bridge
$\tau = RC$	24 μ sec	Time Constant
ϵ_r	1500	Barium Titanate Compound
L	.0254 meter	Film length
w	.0064 meter	Film width
k	1.1×10^{-5} meter	Resistive film thickness
σ	70 mhos/meter	Film conductivity
a	5.1×10^{-4} meter	Substrate thickness

2.2 Instrumentation and Measurement Techniques

The open-circuit transient response of thin-film structures with small time constants cannot be accurately measured with commercially available oscilloscopes. The input impedance of common oscilloscopes is 10 to 50 pf in parallel with 1 to 10 megohms, giving an impedance on the order of only 10 k ohms at 1 Megacycle. This poor approximation of an open circuit distorts the short-time response.

To make the measurements as accurate as possible, a cathode follower was used in the circuit of Figure 8 to provide a high output impedance to the distributed-parameter network. The impedance of the cathode follower, experimentally determined, is a 30 megohm resistor in parallel with a 3-picofarad capacitor.

2.3 Response Time for Resistive Load

Response times for the microsystem terminated in pure resistance appear in Table 6 with the corresponding calculated values, and are plotted in Figure 9.

Table 6. Comparison of Measured and Experimental
Response Times for Resistive Load

R_T (K ohm)	τ_c (microsec.) calculated	τ_c (microsec.) measured	Waveform	Remarks
.57	6.5	6.0	a) Good experimental agreement for all R	
3.82	8.4	7.5		10 to 20 o/o experimental accuracy
5.7	9.0	8.7	b) Approximately 2μsec time delay	
11.4	10.0	10.0		
57	11.5	12.5		

2.4 Response Time for Arbitrary Load

The calculated and measured response times of the microsystem terminated in a series combination of resistance and capacitance are compared graphically in Figure 10. Calculated response times are obtained from the design chart, Figure 7.

2.5 Comparison of Calculated and Observed Waveforms

For large terminating capacitance, the magnitude of the load resistance has a profound effect on the response wave shape; for small load capacitance, the effect is smaller. This is illustrated in Table 7. Note the comparison between predicted and experimental performance.

2.6 Comparison of One-lump and Two-lump Models

As an example, for a source impedance of 0.5 R as shown in Figure 11, the response $v_o(t)$ is calculated from both a one and two-lump model. Results are compared in Table 8. The expressions for $v_o(t)$ were derived by use of standard techniques of network analysis.

The device described in Table 5 was used in the circuit of Figure 12 to determine the response experimentally. In Figure 13 the calculated and experimental responses are compared. Note that an improvement of the one-lump approximation for short-time response would result from the inclusion of a delay line in the model as was suggested in the companion paper.

2.7 Capacitor-Coupled Configurations

As with a transistor, or for that matter any three-terminal device, there are three distinct configurations of a three terminal microsystem structure. These are listed in Table 9. In constrast to the common capacitor, the capacitor-input and capacitor-output circuits exhibit high-pass characteristics; hence, we expect that the one-lump model will not be as accurate for these configurations as for the low-pass common-capacitor configuration. The capacitor-input and capacitor-output configurations are of practical importance since they can serve as capacitive couplers at the input and output of a network.

An open-circuited microsystem in the capacitor-input configuration is driven with a voltage source as shown in Figure 14. The one-lump approximation gives

$$v_o(t) = \exp\left[-\,(2t/\tau)\right]$$

The device described in Table 5, was used in the test setup of Figure 8 to experimentally determine the response. This response, which appear to be purely exponential, had a time constant of 10.5 microseconds which compares favorably with the calculated response time of 12 microseconds.

The previous example should not imply that the one-lump model for capacitive input will give 15 per cent accuracy for arbitrary load and source impedances; in fact, arguments presented in the companion paper suggest that this is unlikely.

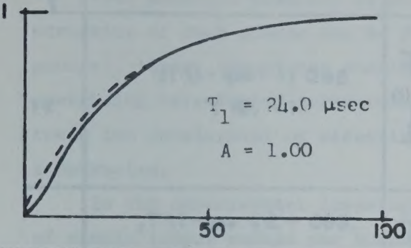
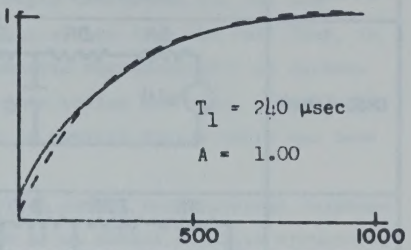
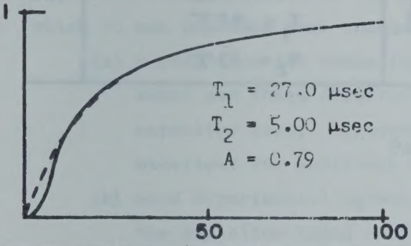
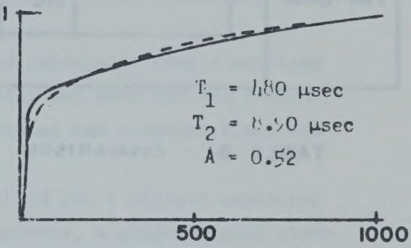
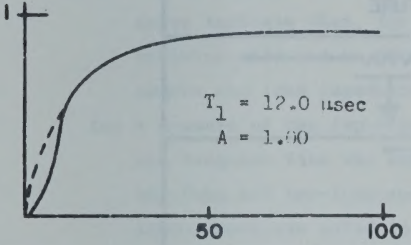
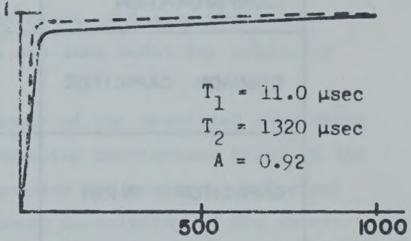
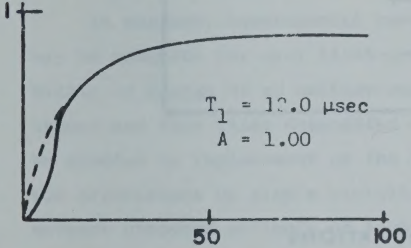
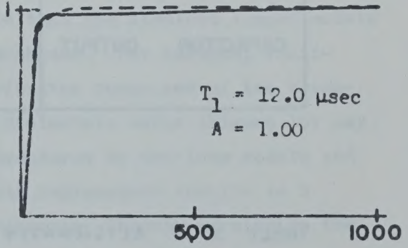
Small Capacitance $C_T/C = 1.0$	R_T in K ohm	Large Capacitance $C_T/C = 10$
 $T_1 = 24.0$ µsec $A = 1.00$	.57	 $T_1 = 240$ µsec $A = 1.00$
 $T_1 = 27.0$ µsec $T_2 = 5.00$ µsec $A = 0.79$	5.7	 $T_1 = 480$ µsec $T_2 = 8.90$ µsec $A = 0.52$
 $T_1 = 12.0$ µsec $A = 1.00$	28.5	 $T_1 = 11.0$ µsec $T_2 = 1320$ µsec $A = 0.92$
 $T_1 = 12.0$ µsec $A = 1.00$	570.	 $T_1 = 12.0$ µsec $A = 1.00$

Table 7

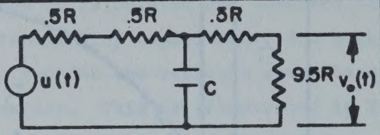
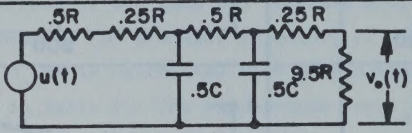
MODEL	EQUIVALENT CIRCUIT	$V_o(t)$	$\frac{T_1}{\mathcal{T}}$
ONE-LUMP	 .5R .5R .5R u(t) C 9.5R $v_o(t)$	$.860\ (1 - \exp - t/T)$ $T = .91\ \mathcal{T}$	.91
TWO-LUMP	 .5R .25R .5 R .25 R u(t) .5C .5C 9.5R $v_o(t)$	$.860 - .99 \exp - t/T_1$ $- .125 \exp - t/T_2$ $T_1 = .82\ \mathcal{T}$ $T_2 = .10\ \mathcal{T}$	.90

TABLE 8. COMPARISON OF MODELS

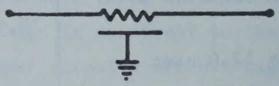
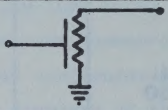
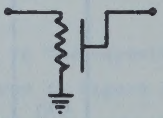
CONFIGURATION	STRUCTURE
COMMON CAPACITOR	
CAPACITOR INPUT	
CAPACITOR OUTPUT	

TABLE 9. ALTERNATIVE CONFIGURATIONS

3. DISCUSSION

The representation of distributed devices by lumped models readily permits the construction of design charts which, for given constraints, graphically display design possibilities.

When moderate accuracy is satisfactory, rapid techniques for the construction of such charts may be developed. This arises from the fact that, in general, linear asymptotes result from a parametric representation of systems containing only lumped parameters, if appropriate scales are used. To illustrate the development of effective techniques, a typical design chart has been constructed.

In the experimental investigations described above, the transient response of simple lumped models was compared with that of an actual thinfilm structure. Different configurations of the device and various circuit environments in which it was embedded were considered, viz.:

(a) A comprehensive comparison of results calculated from a one-lump model and those observed experimentally was made for one common-capacitor case. Agreement of wave shapes and response times was excellent for arbitrary RC load.

(b) Good experimental agreement was obtained for a circuit employing the capacitor-input configuration; however, a comprehensive study was not made. Theoretical arguments presented in the companion paper indicate that, for this configuration, detailed short-time behavior will not be obtained from a one-lump model for arbitrary source and load impedances.

(c) A measure of the improvement in accuracy of the predicted wave shape and response time was obtained by comparing predictions based on the one-lump and two-lump models with observed performance. Principal improvement was noted in a more accurate description of the short-time response.

In summary, experimental results indicate that the simplest lumped models may be adequate for many first-order design problems. For example, feasibility of design of an emitter-coupled multivibrator comprised of two transistors and thin films desposited on a single dielectric wafer (Figure 15) may be studied by replacement of the thin-film structures by one-lump models and the transistors by simple circuit models. This replacement results in a network composed entirely of lumped circuit elements, thereby permitting the use of familiary network techniques. On the other hand, representation of either transistors or thin-film structures by distributed models presents an intractable problem to the circuit designer.

In assessing the results of this investigation, its scope and limitations must be recognized, specifically:

(a) The design chart here developed is to be regarded as only a representative example. Attention has been primarily focussed on a voltage source and a termination consisting of a resistance and a capacitance in series. Clearly the methods proposed can be used to construct design charts applicable for more general termination and drive (e.g., those in Table 4)

(b) Corresponding to the thorough theoretical investigation of the common-capacitor configuration which led to the construction of a design chart for this case, a thorough experimental study of this configuration was made. The capacitor-coupled configurations were treated in a less thorough fashion.

(c) Other thin-film devices may be studied in a manner similar to that presented here and in the companion paper. Three layer devices[5,7] such as the R-C-NR and C-R-NC structures present a possible next step.

(d) A noticeable deficiency of the present study is the lack of an *a priori* test for the adquacy of models. Although a sequence of increasingly accurate solutions for transient problems is available through use of successively higher order lumped models, its usefulness is seriously limited by the lack of a priori estimates on the accuracy of each model. In other words, analytical methods for determining bounds on the accuracy of the lumped models have not been established. Adequacy must be assessed by comparison with experimental results or with the distributed solution.

This deficiency is of paramount importance in assessing the appropriateness of a first-order circuit design for the following reason: If a model is inappropriate, the only analytical recourse consists of successive replacement of the N-lump model by an (n + 1) -lump model until satisfactory convergence in the parameter under consideration (e.g., rise time) results. This approach is time-consuming and tedious. Clearly an *a priori* test is desirable. Although qualitative guidelines for the adequacy of one-lump models were presented in the companion paper, no corresponding quantitative study has been attempted. A study of this type aiming at the prediction of accuracy for an N lump model in an arbitrary circuit environment would be of practical importance.

Several of the above problems are under investigation and will form the subject of a future publication.

REFERENCES

1. F. A. Lindholm and W. W. Happ "Analysis of Transients in R-C Distributed Parameter Networks" (companion paper, this issue) 1963

2. J. Linville, "Lumped Models of Transistors and Diodes", *Proc. IRE*, Vol.46, pp 1141-1152, June 1958.

3. W. W. Happ, "Stray Capacitance in Thin-Film Circuits", Electronic Industries, June 1962.

4. W. D. Fuller and W. W.Happ Design Procedure for Film-Type Distributed Parameter Circuits," Proc. Nat. Electronics Conf., Vol 17, pp 597-610 1961.

5. P. S. Castro and W. W. Happ, "Distributed Parameter Circuit Design Techniques", Proc. Nat. Electronics Conf., Vol. 17, pp 44-70, 1961.

6. W. D. Fuller and W. W. Happ, "Synthesis of Solid State Distributed Parameter Functions", I.R.E. Nat Convention Record, Vol. 10, 1962.

7. W. W. Happ, "Synthesis of Distributed R-C Networks," Proc. I.R.E., pp. 483-484,April 1962.

FIGURE CAPTIONS

Figure 1 through Figure 6. No Captions

Figure 7. Design chart calculated from one-lump model

Figure 8. Instrumentation

Figure 9. Response time vs. resistive load

Figure 10. Experimental and calculated response

Figure 11. Finite source impedance

Figure 12. Circuit to determine waveforms

Figure 13. Comparison of calculated and experimental response

Figure 14. Capacitor-input configuration

Figure 15. Emitter-coupled multivibrator: Monostable operation

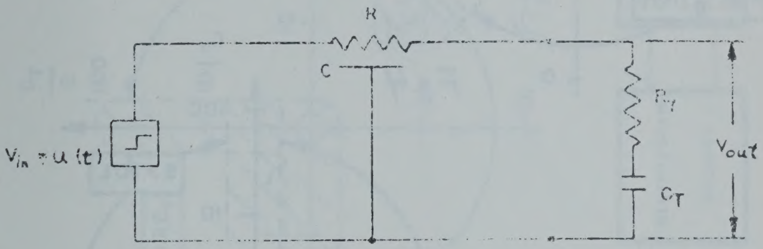


Fig 1

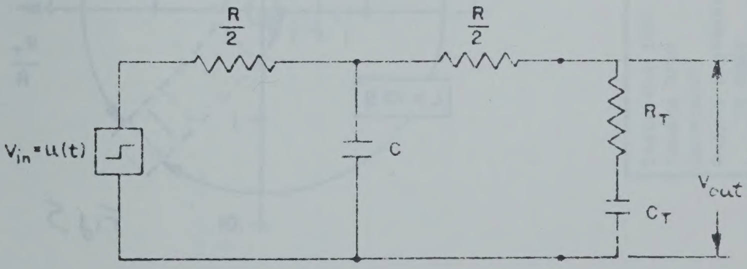


Fig 2

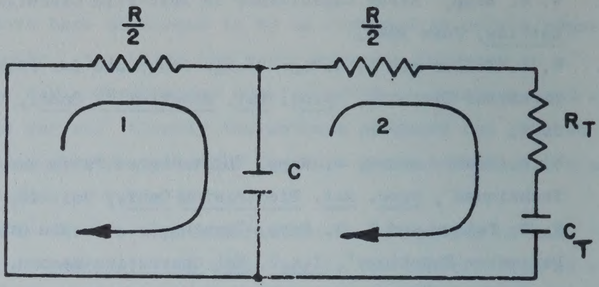
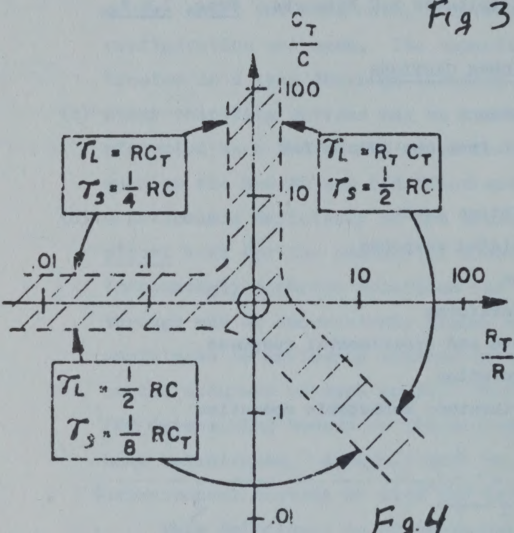

R/2
R/2
1
2
C
R_T
C_T

Fig 3


C_T/C
100
10
.1
.01
.01
.1
10
100
R_T/R
τ_L = RC_T
τ_S = 1/4 RC
τ_L = R_T C_T
τ_S = 1/2 RC
τ_L = 1/2 RC
τ_S = 1/8 RC_T

Fg.4

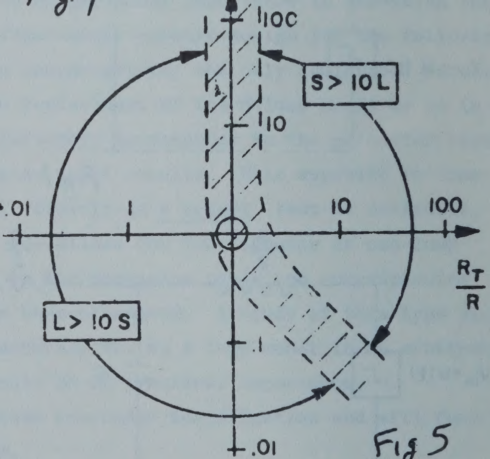

100
10
.01
1
10
100
.1
.01
R_T/R
S > 10 L
L > 10 S

Fig 5

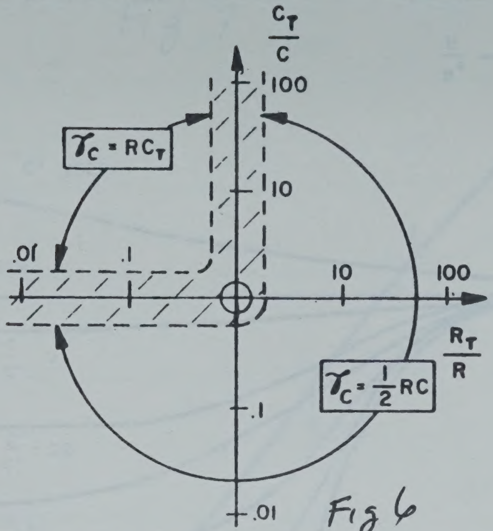

$\frac{C_T}{C}$
100
10
.1
.01
.01
.1
10
100
$\frac{R_T}{R}$
$\tau_c = RC_T$
$\tau_c = \frac{1}{2}RC$

Fig 6

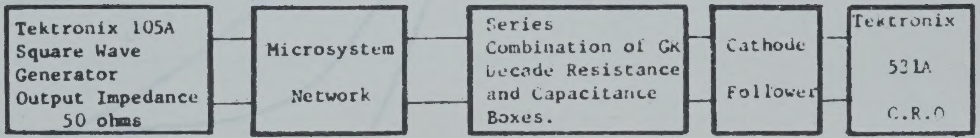

Tektronix 105A Square Wave Generator Output Impedance 50 ohms
Microsystem Network
Series Combination of GR Decade Resistance and Capacitance Boxes.
Cathode Follower
Tektronix 531A C.R.O

Fig 8

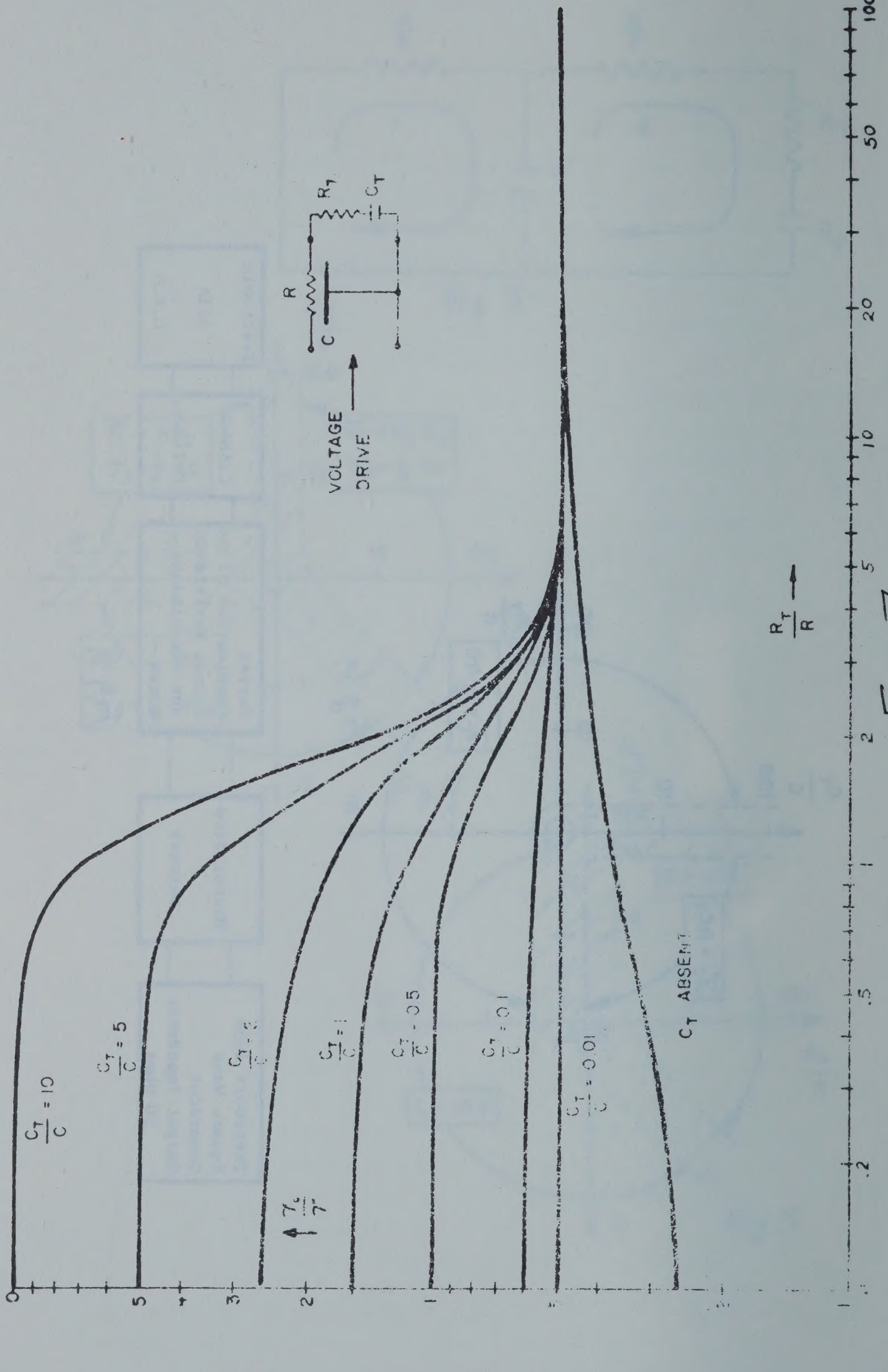

$\frac{C_T}{C} = 10$
$\frac{C_T}{C} = 5$
$\frac{C_T}{C} = 2$
$\frac{C_T}{C} = 1$
$\frac{C_T}{C} = 0.5$
$\frac{C_T}{C} = 0.1$
$\frac{C_T}{C} = 0.01$
C_T ABSENT
$\uparrow \frac{\tau_c}{\tau}$
$\frac{R_T}{R} \rightarrow$
VOLTAGE DRIVE
R
C
R_T
C_T
10
5
4
3
2
1
.1
.2
.5
1
2
5
10
20
50
100

Fig 7

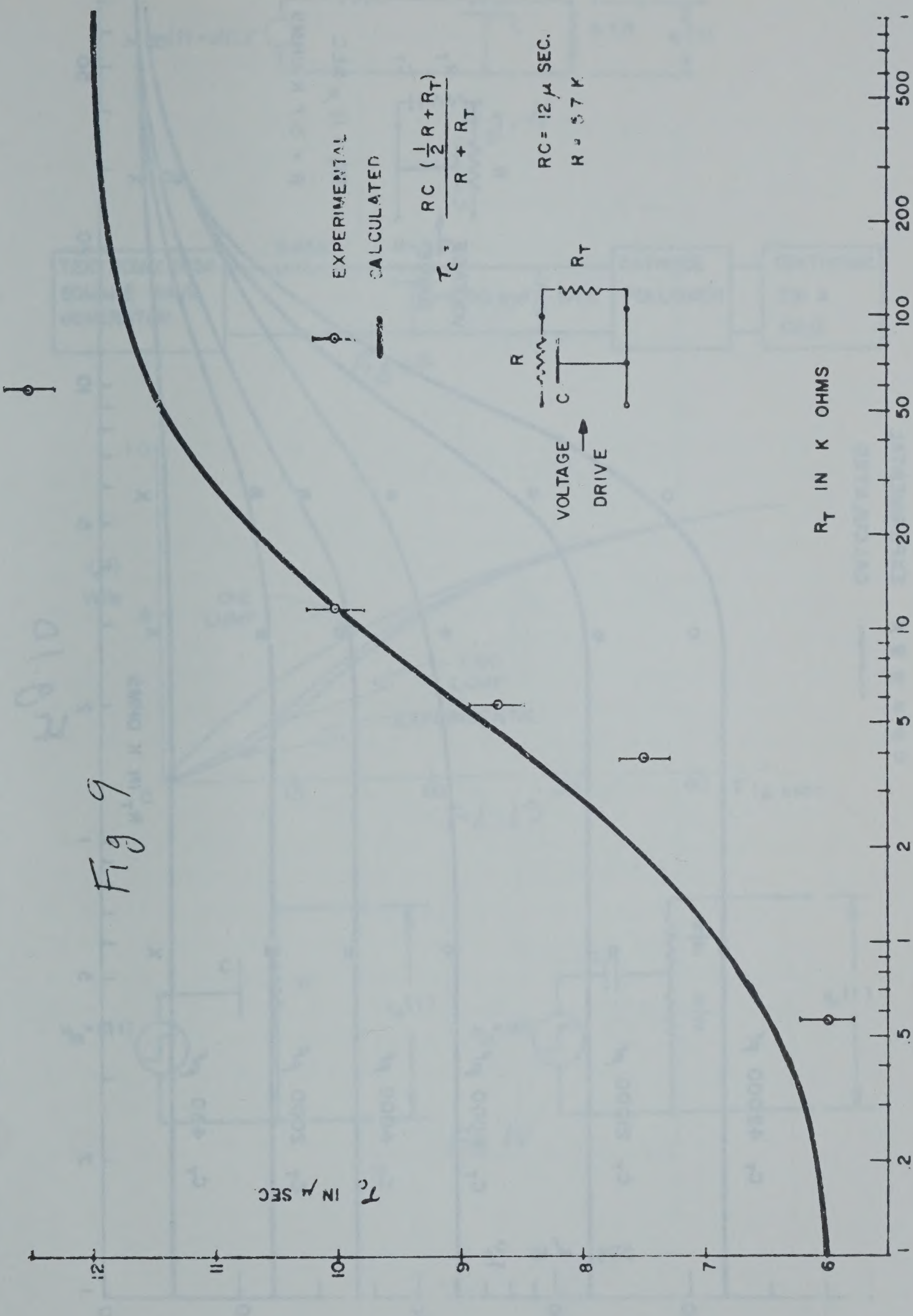

τ_c IN μ SEC
12
11
10
9
8
7
6
.1
.2
.5
1
2
5
10
20
50
100
200
500
1
R_T IN K OHMS
EXPERIMENTAL
CALCULATED
$\tau_c = \frac{RC\ (\frac{1}{2}R + R_T)}{R + R_T}$
RC = 12 μ SEC.
R = 57 K
VOLTAGE DRIVE
R
C
R_T

Fig 9

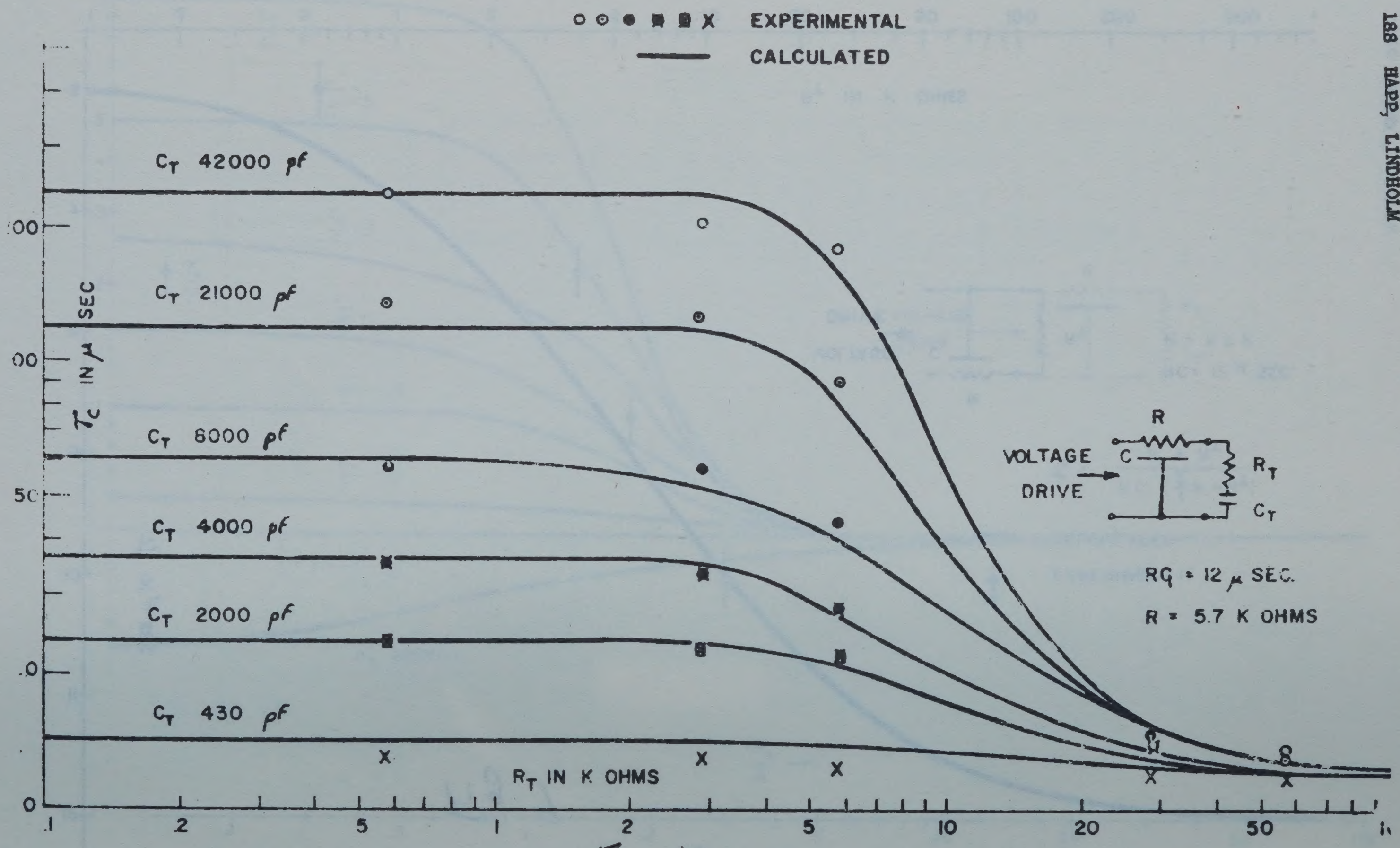

○ ⊙ ● ▪ ▪ X EXPERIMENTAL
CALCULATED
C_T 42000 pf
C_T 21000 pf
C_T 8000 pf
C_T 4000 pf
C_T 2000 pf
C_T 430 pf
τ_c IN μ SEC
00
00
50
.0
0
R_T IN K OHMS
.1
.2
.5
1
2
5
10
20
50
R
VOLTAGE
DRIVE
C
R_T
C_T
RC = 12 μ SEC.
R = 5.7 K OHMS

Fig. 10

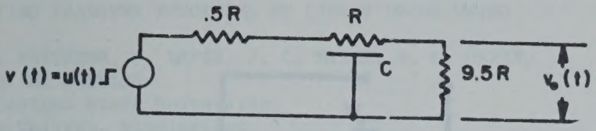

.5 R
R
C
9.5 R
v (t) = u(t)
$v_o(t)$

Fig. 11

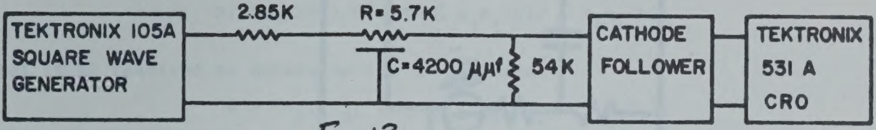

TEKTRONIX 105A SQUARE WAVE GENERATOR
2.85K
R = 5.7K
C = 4200 μμf
54K
CATHODE FOLLOWER
TEKTRONIX 531 A CRO

Fig 12

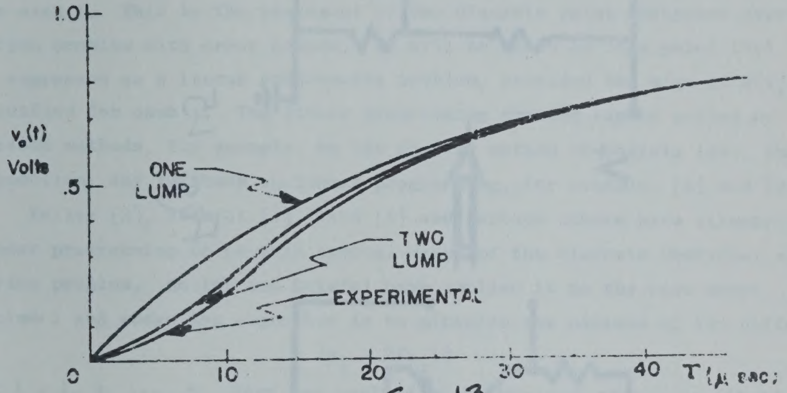


Fig 13

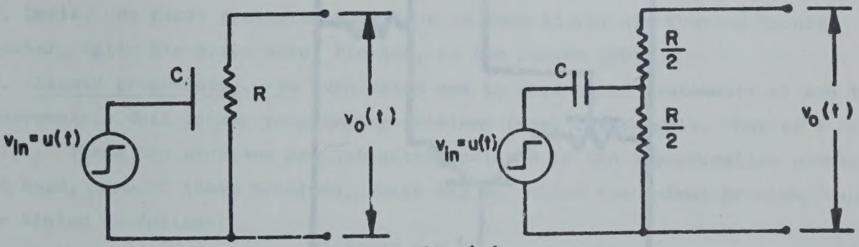

C
R
$v_o(t)$
$v_{in} = u(t)$
C
R/2
R/2
$v_{in} = u(t)$
$v_o(t)$

Fig 14

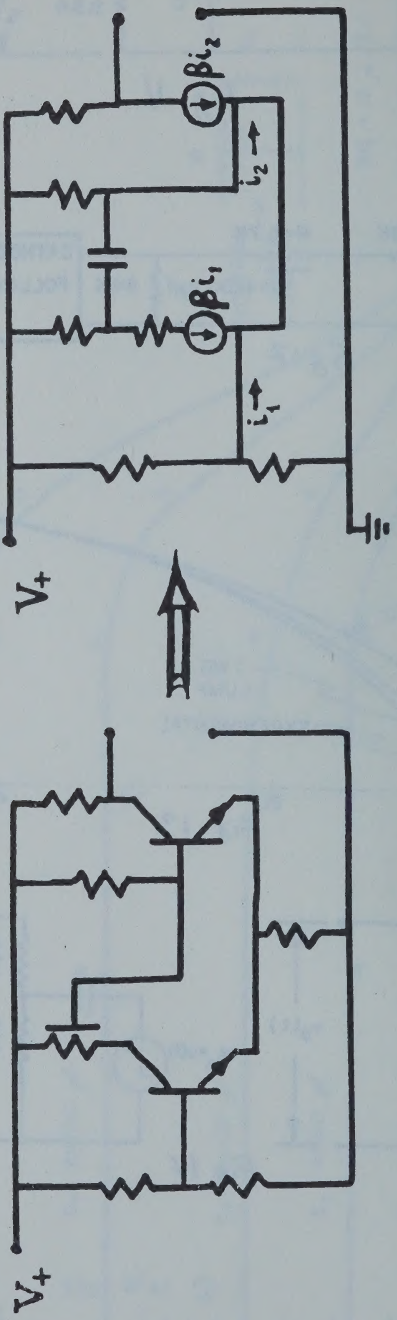

V+
βi₂
i₂
βi₁
i₁
V+

Fig. 15

APPROXIMATING TRANSFER FUNCTIONS BY LINEAR PROGRAMMING

G. A. MATTHEWS, F. DAVIS, J. C. McKEE, R. C. CAVIN,
and G. M. SIBBLES
Mississippi State University
State College, Mississippi

1. Introduction. For i = 1, 2, ..., n, let there be given a set of real, finite pairs (x_i, y_i), a corresponding set of finite quantites $e_i > 0$, and the sets $p_j(x_i)$, j = 1, 2, ..., m, and $q_k(x_i)$, k = 1, 2, ..., t. For

$$R(x) \equiv P(x)/Q(x) \equiv \sum_1^m a_j p_j(x) / \sum_1^t b_k q_k(x),$$

let it be required to determine any (m+t)-tuple

$$c = (a_1, a_2, \ldots, a_m, b_1, b_2, \ldots, b_t),$$

if any exists, such that

$$|y_i - R(x_i)| \leq e_i \tag{1}$$

for each i. This is the statement of the discrete point Chebyshev approximation problem with error bounds. It will be shown in this paper that it can be expressed as a linear programming problem, provided the sign of $Q(x_i)$ is specified for each i. The linear programming problem can be solved by various methods, for example, by the simplex method of Dantzig (see, in this connection, any textbook on linear programming, for example, [1] and [2]).

Kelley [3], Stiefel [4], Ward [5] and perhaps others have already applied linear programming to certain special cases of the discrete Chebyshev approximation problem. Kelley and Stiefel have applied it to the case where $Q(x) \equiv 1$ and where the objective is to minimize the maximum of the differences

$$|y_i - R(x_i)|$$

for i = 1, 2, ..., n. Ward has applied it to the case where, in (1), $Q(x) \equiv 1$. The authors are not aware of its previous application to the more general case considered here. The mathematical developments in this paper are due to F. Davis. He first presented them in a seminar at the Air Proving Ground Center, Eglin Air Force Base, Florida, in the summer 1962.

2. Linear programming. We turn aside now to give brief statements of the two unsymmetric dual linear programming problems (see, for example, Chapter 5 of [1]). These two problems are intimately related to the approximation problem at hand. One of these problems, which may be called the primal problem, can be stated as follows:

$$\text{maximize } g = Wb \tag{2}$$

subject to

$$WA \leq c, \tag{3}$$

where

$$W = (w_1, w_2, \ldots, w_m),$$
$$b^t = (b_1, b_2, \ldots, b_m),$$
$$c = (c_1, c_2, \ldots, c_n)$$

and $A = (a_{ij})$ is a matrix of m rows and n columns. The dual of the primal problem can be stated as follows:

$$\text{minimize } f = cX \tag{4}$$

subject to

$$AX = b, \tag{5}$$

$$X \geq 0, \tag{6}$$

where

$$X^t = (x_1, x_2, \ldots, x_n)$$

and b, c and A are defined in the statement of the primal problem. The solutions of the primal and dual problems can be obtained in the solution of either; further, maximum g = minimum f.

3. The approximation problem as a linear program. We return now to the problem at hand. Refer to the statement of it given in Section 1, define a variable z as indicated below and write, for i = 1, 2, ..., n:

$$\text{minimize } z \tag{7}$$

subject to

$$|y_i - R(x_i)| \leq e_i - z/Q(x_i) \tag{8}$$

for each i for which

$$Q(x_i) > 0, \tag{9}$$

and subject to

$$|y_i - R(x_i)| \leq e_i + z/Q(x_i) \tag{10}$$

for each i for which

$$Q(x_i) < 0. \tag{11}$$

Clearly, $z > 0$ in (7) - (11) implies the (m+t)-tuple

$$c = (a_1, a_2, \ldots, a_m, b_1, b_2, \ldots, b_t)$$

satisfies (1) for each i. On the other hand, $z < 0$ implies c does not satisfy (1) for each i. In fact, it implies that no such c exists, since z is the maximum z consistent with (8) - (11). In short, our problem has a solution for the specified form of R(x) if and only if $z \geq 0$ in (7) - (11).

The system (7)-(11) constitutes a programming problem but it is not linear. In order to obtain an equivalent linear form, let us clear the system of absolute values and fractions. Let us also arrange terms with all unknowns on the left. At the same time, since R(x) is a ratio, which necessitates that we fix one of its coefficients, let $b_1 = 1$. We then obtain, for i = 1, 2, ..., n,

$$\text{maximize } z \tag{12}$$

subject to

$$z + (y_i - e_i) \sum_{2}^{t} b_k q_k(x_i) - \sum_{1}^{m} a_j p_j(x_i) \leq (e_i - y_i)\, q_1\,(x_i), \tag{13}$$

$$z - (y_i + e_i) \sum_{2}^{t} b_k q_k(x_i) + \sum_{1}^{m} a_j p_j(x_i) \leq (e_i + y_i)\, q_1\,(x_i) \tag{14}$$

for each i for which

$$\sum_{1}^{t} b_k q_k (x_i) > 0,$$

and subject to

$$z - (y_i - e_i) \sum_{2}^{t} b_k q_k (x_i) + \sum_{1}^{m} a_j p_j (x_i) \leq (-e_i + y_i)\, q_1\, (x_i), \qquad (15)$$

$$z + (y_i + e_i) \sum_{2}^{t} b_k q_k (x_i) - \sum_{1}^{m} a_j p_j (x_i) \leq (-e_i - y_i)\, q_1\, (x_i) \qquad (16)$$

for each i for which

$$\sum_{1}^{t} b_k q_k (x_i) < 0.$$

In the system (12) - (16), (13) and (14) correspond to (8), and (15) and (16) correspond to (10). Explicit statements of (9) and (11) are not necessary. Relations (13) and (14) imply (9) and relations (15) and (16) imply (11).

The system (12) - (16) constitutes a linear programming problem. It has the form of (2), (3). It can be solved by any of several procedures. It remains now to show that (12) - (16) is equivalent in all essential respects to (7) - (11). We consider the three possible ranges of z, namely, $z > 0$, $z = 0$, $z < 0$.

(a) $z > 0$. When $z > 0$ in (12)-(16), $Q(x_i) \neq 0$ for any i. In fact,

$$z \leq \min |e_i Q(x_i)|$$

for $i = 1, 2, \ldots, n$. Therefore, when $z > 0$, (12) - (16) is equivalent to (7) - (11).

(b) $z = 0$. When $z = 0$ in (12) - (16), the associated c may or may not be meaningful with respect to (7) - (11). It will be meaningful and (12) - (16) and (7) - (11) will be equivalent if and only if $Q(x_i) \neq 0$ for any i. But $Q(x_i) = 0$ may exist for some one or more i. However, when $Q(x_i) = 0$, $P(x_i) = 0$, as can be seen upon study of the pairs (13), (14) and (15), (16). But if $Q(x_i) = 0$ and $P(x_i) = 0$, $Q(x)$ and $P(x)$ have a common factor. Hence, $Q(x)$ and $P(x)$ can be reduced by their common factor and the resultant function will itself be an optimum solution of (12) - (16). The reduction process can be repeated as many times as there are common factors of $Q(x)$ and $P(x)$. When no common factor remains, the resultant function will have no poles at $x = x_i$, $i = 1, 2, \ldots, n$. Further, the resultant function will be an optimum extreme point solution of (12) - (16). As a consequence, its existence will be revealed in the standard solution procedure, after which the solution itself can be easily determined (see chapters 3 and 4 of [1]).

(c) $z < 0$. When $z < 0$ in (12) - (16), there may exist solutions, namely, those for which $Q(x_i) = 0$ for some i, which do not satisfy (7) - (11). But the set of solutions of (8) - (11) is a subset of the solutions of (13) - (16). Hence, $z < 0$ in (12) - (16) implies $z < 0$ in (7) - (11).

It follows from (a), (b) and (c) above that (12) - (16) is equivalent to (7) - (11), in so far as the problem at hand is concerned. Therefore, the Chebyshev approximation problem with error bounds is solved by solving (12) - (16).

It should be noted that in the statement of most practical problems only one of the pairs (8), (9) and (10), (11) will appear in (7) - (11). If only one of these pairs does appear in (7) - (11), then only one of the pairs (13), (14) and (15), (16) will appear in (12) - (16).

If $Q(x_i) \equiv 1$ in (7) - (11), the problem reduces to that solved by Ward [5]. If $Q(x) \equiv 1$ and $e_i = 0$ for all i in (7) - (11), the problem reduces to that solved by Kelley [3] and Stiefel [4].

It is clear that the system (7) - (11) and (12) - (16) can be augmented by restrictions of the form

$$\sum_1^t f_k(b_k) + \sum_1^m g_j(a_j) = 0.$$

Restrictions of this type are often necessary in practical applications.

4. The computational problem. Refer to (12) - (16). These constraints form a system of 2n linear inequalities in m+t unknowns. In the conventional solution approach, the inequalities must be transformed to equivalent equalities by adding a slack variable to each inequality. At the same time, each original unknown must be replaced by the difference of two unknowns to allow freedom for negative unknowns to appear in the solution. The resultant transformed system consists of 2n linear equations in 2n + 2m + 2t unknowns. This is a formidable system to solve even for relatively small n. As was pointed out in Section 2, the solution of a linear programming problem can be accomplished by solution of its dual. The dual of (12) - (16) has the form of (4) - (6). After artificial variables have been added, the dual of (12) - (16) consists of m + t equations in 2n + m + t unknowns. This system is smaller than the primal system. It is sufficiently small for most realistic problems that it can be handled quite well even by small computers.

5. Magnitude-squared transfer functions. In this section we shall illustrate the application of the linear programming method to several types of magnitude-squared transfer functions. For all types the basic problem is the following: given a set of specifications for a voltage transfer ratio, G, in terms of the angular frequency, w, to find a ratio of functions expressed as a function of the complex frequency, s, which satisfies the specifications and meets all the network realizability requirements.

The first step in obtaining the required

$$G(s) = P(s)/Q(s)$$

is to obtain a magnitude-squared function

$$|G^2(jw)| = P(w^2)/Q(w^2)$$

which satisfies the specifications. To fix the idea, let $Q(w^2)$ be positive for all w. Then $G^2(jw)$ must satisfy the following (and yield $z > 0$):

$$\text{maximize } z \tag{17}$$

subject to

$$\left| |G^2(jw_i)| - P(w_i^2)/Q(w_i^2) \right| \le e_i - z/Q(w_i^2), \tag{18}$$

where e_i, for $i = 1, 2, \ldots, n$, is the allowable error at the point $(w_i, |G^2(jw_i)|)$.

In addition to the constraints (18), others of the form

$$\sum_1^t f_k(b_k) + \sum_1^m g_j(a_j) \le 0, \tag{19}$$

where f and g are linear, can be specified. This permits considerable freedom in the definition of $|G^2(jw)|$. For example,

$$|G^2(jw_i)| = P(w_i^2)/Q(w_i^2)$$

can be specified for any point $(w_i, |G^2(jw_i)|)$. Other examples of the application of (19) are

$$r_i \le a_i \le s_i,$$

$$a_i \le k_j a_j.$$

The system (17) - (19) can be related to a linear programming problem in the manner described in Section 3.

If $z \geq 0$ in the solution of (17) - (19), a plot of $|G^2(jw)|$ will indicate whether $|G^2(jw)|$ satisfies the specifications. Obviously, the specifications are satisfied at $(w_i, |G^2(jw_i)|)$. Experience in solving many problems of the type considered here indicates a high probability that they are satisfied everywhere. If they are not, simple applications of constraints of the type (19) will usually suffice to make them satisfied everywhere.

Example 1 - the low pass filter. In order to show the flexibility of the linear programming method, let us require that $|G(jw)|$ approach the upper bound in the passband and the lower bound in the stopband. Additionally, require that

$$|G(j1.2)| = .707.$$

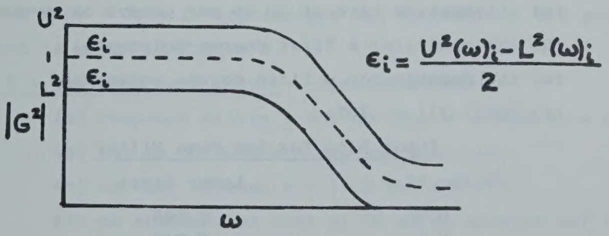


Figure 1.

Refer to Figure 1. For arbitrarily chosen points, $(w_i, |G^2(jw_i)|)$, $i = 1, 2, \ldots, n$, and specified forms of $P(w^2)$ and $Q(w^2)$ $Q(w^2)>0$,state the following:

$$\text{maximize } z \tag{20}$$

subject to

$$|U^2(jw_i)| - P(w_i^2)/Q(w_i^2) \leq 2e_i - z/Q(w_i^2), \tag{21}$$

$$-|U^2(jw_i)| + P(w_i^2)/Q(w_i^2) \leq 0 \tag{22}$$

in the passband,

$$-|L^2(jw_i)| + P(w_i^2)/Q(w_i^2) \leq 2e_i - z/Q(w_i^2), \tag{23}$$

$$|L^2(jw_i)| - P(w_i^2)/Q(w_i^2) \leq 0 \tag{24}$$

in the stopband, and

$$P((1.2)^2)/Q((1.2)^2) \leq (.707)^2 \tag{25}$$

$$-P((1.2)^2)/Q((1.2)^2) \leq -(.707)^2. \tag{26}$$

As pointed out earlier, relations of the form

$$\sum_1^t f_k(b_k) + \sum_1^m g_j(a_j) \leq 0 \tag{27}$$

may be combined with (21) - (26).

Since z in (21) - (27) is maximized, it will not be negative if non-negative values exist. For non-negative values of z, (21) prevents $|G^2(jw)|$ from taking values below the lower bound, and (22) prevents it from exceeding the upper bound. In the stopband, (23) and (24) have the effects of (22) and (21), respectively. Statements (25) and (26) force the equality $|G(j1.2)| = .707$, and (27) restricts the a_j and b_k as desired. Finally, (21) tends to force $|G^2(jw)|$ toward the upper bound in the passband, and (23) tends to force it toward the lower bound in the stopband.

The system (20) - (27) constitutes a general form for solution of the low pass filter network approximation problem.

A particular problem. Specifications:

(a) cut off frequency at w = 1.2,

(b) response within ± 0.5 db for $0 \leq w \leq 1$,

(c) attenuation rate of 30 db per octave or greater at cut off,

(d) the numerator a first degree polynomial,

(e) the denominator a fifth degree polynomial,

(f) $|G(j1.2)| = .707$.

Input Data for Low Pass Filter

Data Point	Value of w	Lower Limit	Upper Limit
1	0.0	0.945	1.05
2	0.1	0.945	1.05
3	0.2	0.945	1.05
4	0.3	0.945	1.05

Data Point	Value of w	Lower Limit	Upper Limit
5	0.4	0.945	1.05
6	0.5	0.945	1.05
7	0.6	0.945	1.05
8	0.7	0.945	1.05
9	0.75	0.945	1.05
10	0.8	0.945	1.05
11	0.825	0.945	1.05
12	0.85	0.945	1.05
13	0.875	0.945	1.05
14	0.9	0.945	1.05
15	0.925	0.945	1.05
16	0.95	0.945	1.05
17	0.975	0.945	1.05
18	1.0	0.945	1.05
19	1.01	0.707	1.05
20	1.1	0.707	1.05
21	1.2	0.707	0.707
22	1.3	0.00	0.45
23	1.4	0.00	0.29
24	1.5	0.00	0.2
25	1.6	0.00	0.14
26	1.7	0.00	0.1
27	1.8	0.00	0.07
28	1.9	0.00	0.05
29	2.0	0.00	0.04

The solution to magnitude-squared data is:

$$|G^2(jw)| = \frac{1.1025 + 1.429w^2}{1 + 1.6003w^2 - 4.052w^4 + 16.772w^6 - 22.17w^8 + 9.152w^{10}}$$

The roots in terms of the complex frequency variable s which are located in the left half plane are found to be:

Zeros: $s = -0.87822$

Poles: $s = -0.49869$

$s = -0.44476 \pm j0.56868$

$s = -0.17578 \pm j1.1139$

These are now combined and the final voltage transfer ratio becomes:

$$G(s) = \frac{0.395s + 0.347}{s^5 + 1.739s^4 + 2.724s^3 + 2.364s^2 + 1.318s + 0.331}$$

The system (20) - (27) is appropriate for solution of the problem. A plot of the solution is given in Figure 2.

Example 2 - the band pass filter. Specifications:

(a) response within ± 0.5 db from w = 0.85 to w = 1.15,

(b) cut on between w = 0.8 and w = 0.85,

(c) cut off between w = 1.15 and w = 1.2,

(d) an attenuation rate of 30 db or greater per octave at cut on and cut off,

(e) highest order term in numerator is 4,

(f) higherst order term in denominator is 5.

The system (17) - (19) is appropriate for solution of the problem. A plot of a solution based on 30 typical data points is given in Figure 3.

Example 3 - the high pass filter. Specifications:

(a) response within ± 0.5 db from w = 1.0 to w = 2.0,

(b) cut on between w = 0.8 and w = 1.0,

(c) an attenuation rate of 30 db or greater per octave at cut on,

(d) highest order term in numerator is 4,

(e) highest order term in denominator is 4.

The system (17) - (19) is appropriate for solution of the problem. A plot of a solution based on 30 typical data points is given in Figure 4.

6. Fitting functions to both magnitude and phase angle specifications. Let (w_i, v_i), $i = 1, 2, \ldots, k$, be k points, w_i real and $v_i = y_i + jy_i'$. Let $P(jw)$ and $Q(jw)$ be polynomials with real coefficients, $a_1, a_2, \ldots, a_m$ and $b_1, b_2, \ldots, b_t$, respectively. Define $r(jw_i)$ such that

$$r(jw_i) = y_i + jy_i' - P(jw_i)/Q(jw_i) \tag{28}$$

for $i = 1, 2, \ldots, k$. Multiply (28) by $Q(jw_i)$ and let

$$e_i + je_i' = r(jw_i)Q(jw_i).$$

Then

$$e_i + je_i' = y_iQ(jw_i) + jy_i'\,Q(jw_i) - P(jw_i) \tag{29}$$

for $i = 1, 2, \ldots, k$. Let

$$P(jw_i) = P_m(w_i) + jP_n(w_i) \tag{30.1}$$

and

$$Q(jw_i) = Q_m(w_i) + jQ_n(w_i). \tag{30.2}$$

Substituting for $P(jw_i)$ and $Q(jw_i)$ from (30) into (29) and separating real and imaginary parts yields

$$e_i = y_iQ_m(w_i) - y_i'Q_n(w_i) - P_m(w_i), \tag{31}$$

and

$$e_i' = y_i'Q_m(w_i) + y_iQ_n(w_i) - P_n(w_i),\ i = 1, 2, \ldots, k. \tag{32}$$

Equations (31) and (32) can be put into a form suitable for linear programming solution by letting

$$z = \max |e_i, e_i'|\ ,\ i = 1, 2, \ldots, k.$$

The linear programming statement of the problem is as follows:

$$\text{maximize } z \tag{33}$$

subject to

$$z + y_iQ_m(w_i) - y_i'Q_n(w_i) - P_m(w_i) \leq 0, \tag{34}$$

$$z - y_i Q_m(w_i) + y_i' Q_n(w_i) + P_m(w_i) \le 0, \qquad (35)$$

$$z + y_i' Q_m(w_i) + y_i Q_n(w_i) - P_n(w_i) \le 0, \qquad (36)$$

$$z - y_i' Q_m(w_i) - y_i Q_n(w_i) + P_n(w_i) \le 0, \qquad (37)$$

for i = 1, 2, ..., k. As pointed out earlier, at least one of the coefficients in (34) - (37) must be fixed.

In addition to (34) - (37), other constraints of the form

$$\sum_1^t f_k(b_k) + \sum_1^m g_j(a_j) \le 0, \qquad (38)$$

where f and g are linear, can be imposed.

In the system (33) - (38), -z cannot be less than the largest element in a set of positive quantities. Hence, z cannot be positive. The upper bound of z is zero, and when we maximize z we minimize the maximum error in (31) and (32).

The following examples will indicate how linear programming in the form of the system (33) - (38) can be applied to fitting functions to both magnitude and phase angle specifications.

Example 1 - ideal case. An ideal delay line could have the following specifications (where they apply only over the range indicated):

(a) w, the normalized frequency, has the range $0 \le w \le 9$,

(b) the time delay is a constant, $T_d = 1$,

(c) the magnitude of the transfer function equals one,

(d) the maximum degree of the denominator is 6 (the degree of the numerator is the same as that of the denominator for "all-pass" functions), and

(e) the error in time delay is to be a minimum.

The time delay, T_d, is related to the phase shift, B, by

$$T_d = dB/dw. \qquad (39)$$

From the specifications (a) - (e) and (39), B has the value

$$B = w. \qquad (40)$$

The required input data may then be found from (40) and (b). They are given in the following table, where R and X are the real and imaginary components of G_{12}, respectively.

Data Point	Value of w	R	X
1	.0000	1.0000	.0000
2	.5000	.8776	-.4794
3	1.0000	.5403	-.8415
.	.	.	.
.	.	.	.
.	.	.	.
19	9.0000	-.9111	-.4121

The resulting "all-pass" transfer function is

$$G_{12}(s) = \frac{(s - 5.5910 \pm j2.4325)(s - 5.4299 \pm j7.0236)(s - 4.1414 \pm j10.953).}{(s + 5.5910 \pm j2.4325)(s + 5.4299 \pm j7.0236)(s - 4.1414 \pm j10.953)} \tag{41}$$

The maximum magnitude error is .0000. The maximum phase error, which is related to T_d by (39), is .0942 degrees. The errors in this example and in the following one were determined at relatively small intervals on w.

The function (41) was determined by formulating a linear programming problem based on the model (33) - (38).

Example 2 - delay equalizer. The following specifications are actual requirements for a television circuit (f in megacycles):

(a) the envelope of delay of the equalizer is to be constant, $T_d = 1$ microsecond, over the frequency interval, $0 \le f \le 3$,
(b) the envelope of delay of the equalizer is to drop linearly in the frequency interval, $3 \le f \le 4.25$, so that the delay at 3.58 megacycles is $T_d = .83$ microsecond,
(c) the maximum deviation from the characteristic is to be at most ± .02 microseconds on the complete frequency interval, $0 \le f \le 4.25$ (this is a maximum phase error of ± 3.04 degrees),
(d) the transfer function is to be "all-pass",
(e) the maximum degree of the denominator is to be 3.

From the specifications (a) - (e), where the normalized value of w is 1 for each megacycle,

$$T_d = dB/dw = 1 \tag{42}$$

for $0 \le w \le 3$, and

$$T_d = dB/dw = -.293w + 1.88 \tag{43}$$

for $3 \le w \le 4.25$. From (d)

$$B = w$$

for $0 \le w \le 3$, and

$$B = -.146\, w^2 + 1.88w - 1.32$$

for $3 \le w \le 4.25$. The required input data is found from (42) and (43). The resulting "all-pass" transfer function is

$$G_{12}(s) = \frac{-.004197s^3 + 1.009s^2 - .4833s + 1.0000.}{.004197s^3 + 1.009s^2 + .4833s + 1.0000} \tag{44}$$

The maximum magnitude error is .0000 and the maximum phase shift error is ± 1.19 degrees.

The function (44) was determined by solving a linear programming problem based on the form (33) - (38).

7. Summary. What might appear to be a formidable computational problem resolves itself into the simple problem of stating the conditions that a transfer function must satisfy. The basic idea is simply that if a curve can be repre-

sented by a set of points, then a function can be found to approximate these points to any desired degree of accuracy. The user has the freedom to specify what error he will tolerate and insist that the function satisfy the specified error limits; or he can determine a function without specifying the error and then decide whether or not the function is satisfactory.

In general, the only data required are

(a) the degree of P and Q,
(b) a set of data points,
(c) the specified error at each point (optional),
(d) coefficient restrictions (optional).

The preparation of data is the only work in which the user would be involved, for linear programming programs are readily available and programs which will prepare data for input to the linear programming programs are easily written.

The ability to control coefficients is one of the strong points of this method of obtaining transfer functions. Physical realizability can be guaranteed and the form of the function can be controlled. This can simplify the synthesis problem.

In the example given, $G(s) = HP(s)/Q(s)$, the requirement that $0 \leq H \leq \min Q(s)/P(s)$ is not met. Additional restrictions or the addition of an amplifier could easily fulfill this requirement.

REFERENCES

1. Saul Gass, Linear Programming, McGraw-Hill, New York, 1958.
2. S. Vajda, Mathematical Programming, Addison-Wesley, Reading, Massachusetts, 1961.
3. James E. Kelley, "An Application of Linear Programming to Curve Fitting", J. Soc. Indust. Appl. Math., vol. 6, 1958, pp. 15-22.
4. E. Stiefel, "Note on Jordan Elimination, Linear Programming and Tchebycheff Approximation", Numer. Math., vol. 37, pp. 1-17, 1960.
5. L. E. Ward, Jr., "Linear Programming and Approximation Problems", Amer. Math. Monthly, vol. 62, no. 1, pp. 46-53, 1961.

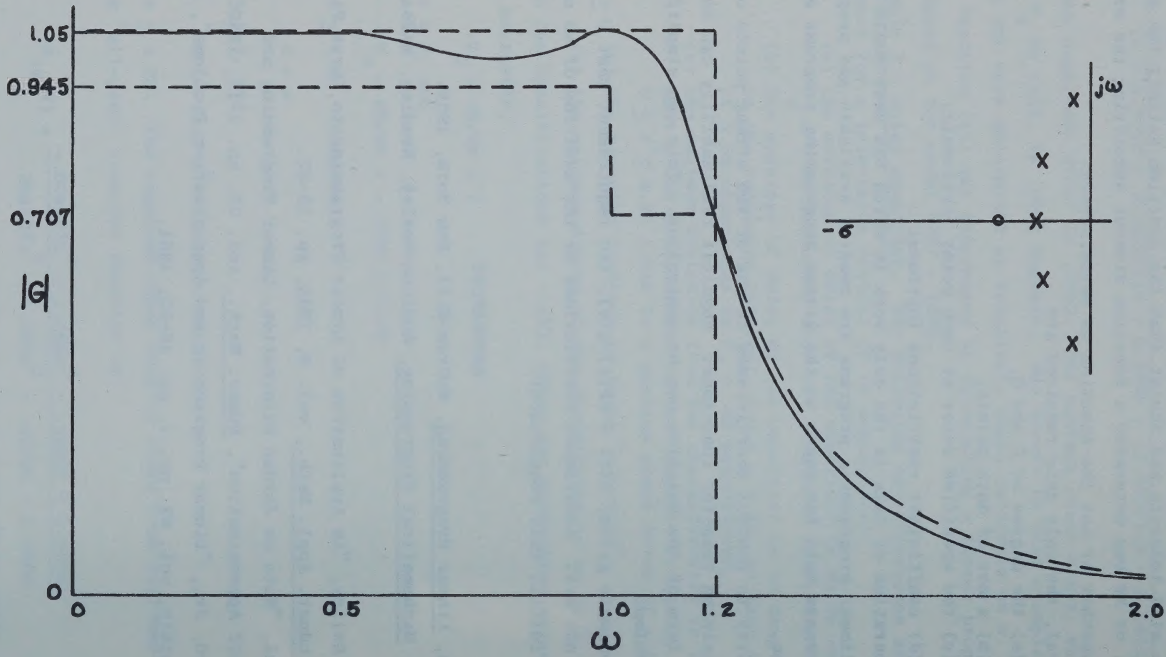
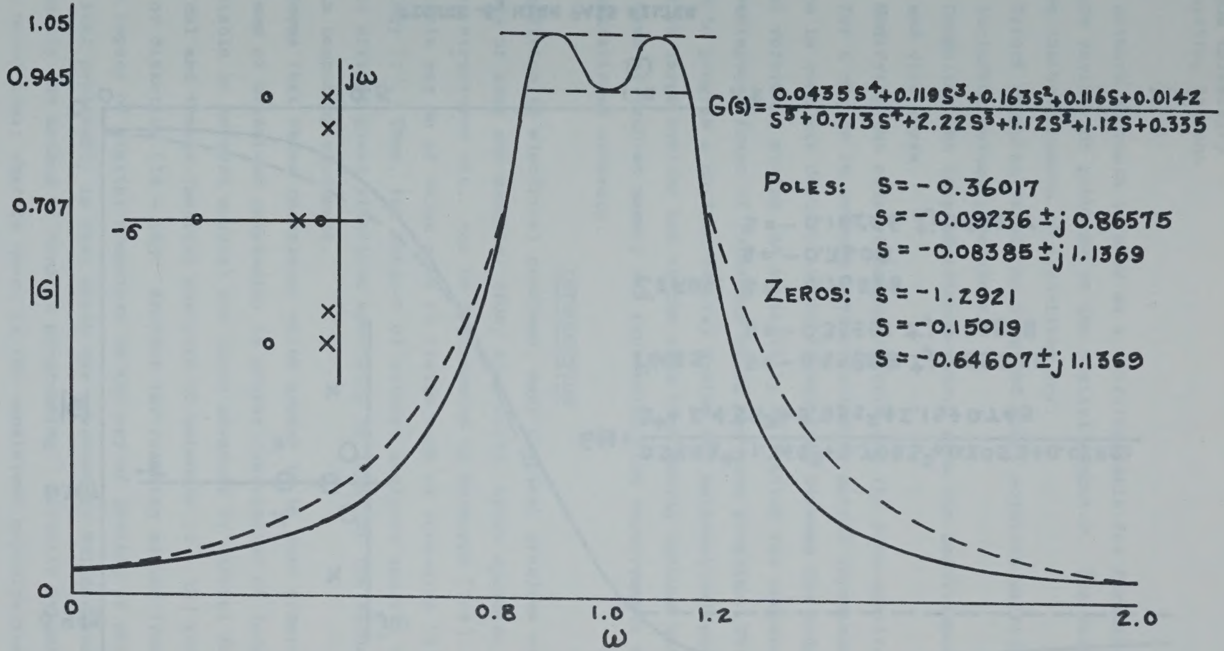


FIGURE 2, LOW PASS FILTER



$$G(s)=\frac{0.0435s^4+0.119s^3+0.163s^2+0.116s+0.0142}{s^5+0.713s^4+2.22s^3+1.12s^2+1.12s+0.335}$$

Poles: $s=-0.36017$
$s=-0.09236\pm j\,0.86575$
$s=-0.08385\pm j\,1.1369$

Zeros: $s=-1.2921$
$s=-0.15019$
$s=-0.64607\pm j\,1.1369$

FIGURE 3, BAND PASS FILTER

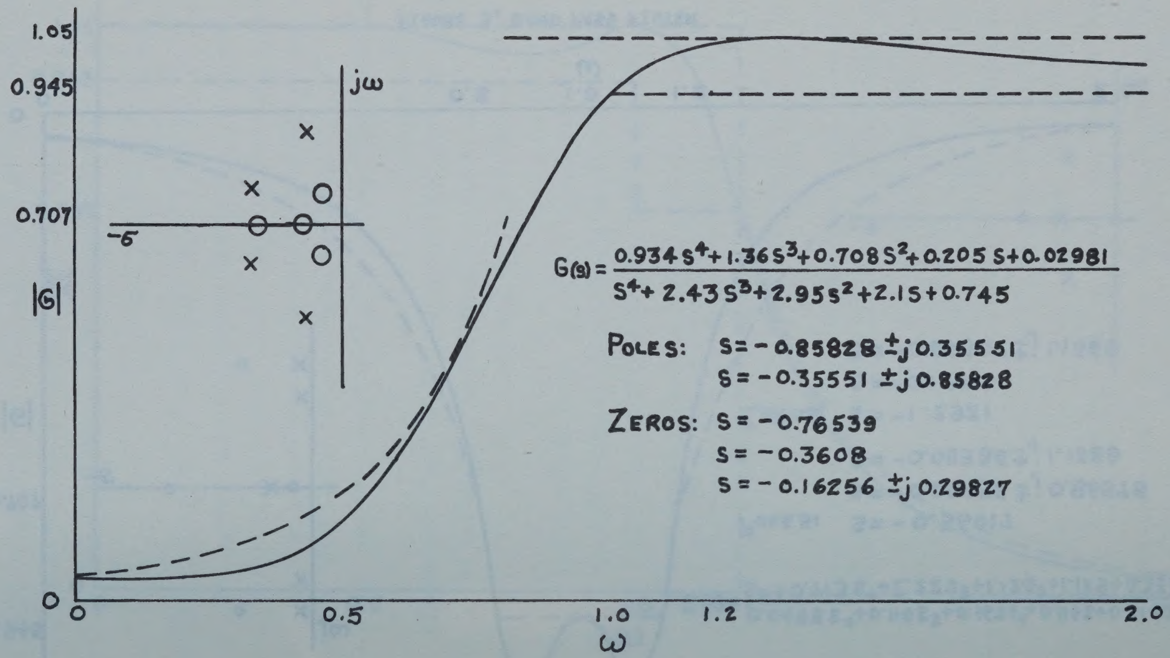


FIGURE 4, HIGH PASS FILTER

DIGITAL COMPUTER ANALYSIS OF LARGE LINEAR SYSTEMS

A. H. EL-ABIAD
Purdue University
Lafayette, Indiana

SUMMARY

The network approach is used as a unifying basis for formulating and solving a wide variety of problems on the digital computer. The basic contributions are the development of algorithms for:

1. Direct computation of the loop and nodal solution matrices of networks including mutual impedances.
2. Computing the loop-admittance-matrix from the nodal-impedance-matrix and vice versa.
3. Modifying the nodal-impedance-matrix and the loop-admittance-matrix for a change in network matrix including mutual impedance.

There is complete duality in the formulation between the nodal and loop frames of reference giving the possibility of letting the computer choose the most advantageous frame of reference using the same program. This method makes possible to program a unified digital system for mathematical modelling and analysis of large systems and at the same time making optimum use of the computer in terms of reduced memory and processing time requirements and maintaining high computational accuracy.

INTRODUCTION

In addition to electrical problems, many physical problems originating from the fields of heat and material flow, elasticity, hydro dynamics, polyatomic molecular structures etc., can be represented by networks [1-4]. Network or graph models may be of value even in fields such as economics [5] sociology[6] and biology [7]. Thus, the subject of network analysis and its automation continues to attract great attention specially in light of the technological advances in computing equipment.

It seems that those techniques which appear to be most promising in studying systems of organized complexity (a proper description of large-scale systems representable by network models) are those advanced by Gabriel Kron; namely topological and tensor (matrix) analysis of networks [8 - 12] and the method of tearing or diakotics [13 - 18]. Another far reaching method (resulting primary from the impact of digital computers on the way of looking at physical and mathematical problems), is that which was advanced by Bellman and his collaborators; namely the method of dynamic programming. Actually dynamic programming is not a novel idea; what is novel is the consistent exploitation of the point of view, which obliges one to reformulate most problems in terms of a property X which is recursive, and the construction of a functional equation which preserves the invariance of X [19]. Sometimes the name invariant imbedding is being used to describe this method. The main advantage is instead of representing a complex system as a multi-dimensional process, it is represented as a sequence of much lower dimensional processes.

Figure 1 summarizes procedures of derivation of the main network matrices by classical methods based on linear transformations and matrix inversion. The alternative procedures, which combines the above mentioned three points of view, and which is presented in this paper is summarized in figure 2. These procedures take advantage of the fact that in almost all engineering systems, the individual elements or components have little or no mutual interaction before they are connected together into the systems structure (network). Because of space limitation, algorithms and equations are presented without proofs which can easily be done by applying Woodbury's formula [20] or circuit theory [21].

DEFINITIONS AND NOTATION

A matrix is denoted by a base letter with or without indices and enclosure according to the following notations:

Base Letters

1. Lower-case Latin letters indicate quantities belonging to the branch frame of reference independent of any connections into a network.
2. Upper-case Latin letters indicate quantities belonging to one of the reference frames which depend on the network connection.

Indices

1. Fixed indices are denoted by lower-case Latin letters while variable indices are denoted by lower-case Greek letters.
2. Letters at the beginning of the alphabet (Latin or Greek) denote loops and letters at the middle denote buses.

Enclosures

1. Square brackets employed around a base letter without indices or with fixed indices will denote a matrix whose order would be obvious from the base letter and/or fixed indices.
2. A base letter with variable indices is always a matrix and does not need brackets, for example:
   - A base letter with two fixed indices is a scalar.
   - A base letter with one fixed and one variable index is a matrix of order one, i.e., a column or a row vector.
   - A base letter with two variable indices is a matrix of order two and is equivalent to a base letter with two fixed indices and square brackets.

Branch-Impedance-Matrix and Branch-Admittance-Matrix

The branch-impedance-matrix consists of the branch (element) self and mutual impedances independent of any network connection and is denoted as follows:

z_{branch} in general

$z_{\mu-\nu,\ \rho-\sigma} = \left[z_{p-q,\ r-s} \right]$ branches identified by buses

$z_{\pm\alpha\pm\beta\pm\ \ldots,\ \pm\gamma\pm\delta\pm\ \ldots} = \left[z_{\pm a\pm b\pm\ \ldots,\ \pm c\pm d\pm\ \ldots}\right]$ branches identified by loops

The branch-admittance-matrix is the inverse of the branch-impedance-matrix and is denoted in the same way with y replacing z. This inversion is very easy since the branch-impedance-matrix is generally of a diagonal form.

Loop-Impedance-Matrix and Loop-Admittance-Matrix

The loop-impedance-matrix refers to a derived matrix involving as many loops as there are independent ones in the network and is denoted as follows:

Z_{loop} in general

$Z_{\alpha,\beta} = \left[Z_{a,b}\right]$ loops denoted by letters at the beginning of the alphabet

The loop-admittance-matrix is the inverse of the loop-impedance-matrix and is denoted in the same way with Y replacing Z.

Node-Pair-Admittance-Matrix and Node-Pair-Impedance-Matrix

The node-pair-admittance-matrix refers to a derived matrix involving as many node-pairs as there are independent ones in the network, such as the set that is made of the terminals of the tree branches and which is denoted as follows:

$Y_{\text{tree n.p.}}$ in general

$Y_{\mu-\nu,\ \rho-\sigma} = \left[Y_{p-q,\ r-s}\right]$ node-pairs denoted by tree branch bus terminals

The node-pair-impedance-matrix is the inverse of the node-pair-admittance-matrix and is denoted in the same way with Z replacing Y.

Bus-Admittance-Matrix and Bus-Impedance-Matrix

The same as the previous case except that the node-pair frame of reference is chosen with one common reference bus and are denoted as follows:

Z_{bus} or Y_{bus} in general

$\left[Z_{m,n}\right]$ or $\left[Y_{m,n}\right]$ denoted by buses assuming that the common reference is known.

Network-Impedance-Matrix and Network-Admittance-Matrix

These refer to the network and have as many rows and columns as there are branches. The elements of the network-impedance-matrix and the elements of the network-admittance matrix do not have mutually inverse relations, but they are related according to equations (12) and (15) which are given later. They are defined as follows:

$Z_{\text{network}} = Z_{\mu-\nu,\ \rho-\sigma} = \left[Z_{p-q,\ r-s}\right]$

where $Z_{p-q,r-s}$ = voltage of p over q due to one p.u. current applied across r-s from r to s. p-q & r-s any two branches in the network.

$$Y_{network} = Y_{\mu-\nu,\rho-\sigma} = \left[Y_{p-q,r-s} \right]$$

where $Y_{p-q,r-s}$ = current in branch p-q from p to q due to one p.u. voltage inserted in series with branch r-s in direction r to s.

ALGORITHM TO COMPUTE Z_{bus}

This algorithm was previously published by the author [21] and is being repeated here in new condensed notation.

The bus-impedance-matrix is calculated by constructing the network, one element at a time, and at the same time carrying out the required operations to obtain the modified matrix as each subsequent element is added. Each newly added element p-q must have at least its terminal p already included in the previously assembled subnetwork.

1. Adding a branch p-q, which is coupled to branches ρ-σ of the previously assembled subnetwork and which adds a new bus q not previously included in the matrix, adds a new row and column to it, the elements of which are calculated as follows:

$$Z_{n,q} = Z_{n,p} + \frac{y_{p-q,\ \rho-\sigma}\, Z_{\rho-\sigma,n}}{y_{p-q,\ p-q}} \tag{1}$$

and

$$Z_{q,q} = Z_{p,q} + \frac{1 + y_{p-q,\ \rho-\sigma}\, Z_{\rho-\sigma,q}}{y_{p-q,p-q}} \tag{2}$$

where

$Z_{n,p}$ = element of the previous bus-impedance matrix

$y_{p-q,p-q}$ & $y_{p-q,\rho-\sigma}$ = elements of the assembled branch-admittance matrix, where it is subdivided as follows

$$y_{branch} = \begin{array}{c|c|c} & p-q & \rho-\sigma \\ \hline p-q & y_{p-q,p-q} & y_{p-q,\rho-\sigma} \\ \hline \rho-\sigma & & \end{array}$$

$Z_{\rho-\sigma,n}$ = column matrix whose elements are computed from the previous bus-impedance-matrix according to: $Z_{\rho-\sigma,n} = Z_{\rho,n} - Z_{\sigma,n}$
n denotes any bus of the previously assembled subnetwork.

2. Adding branch p-q which is coupled to branches ρ-σ of the previously assembled subnetwork and which closes a loop L of the network adds a new row and column whose elements are calculated as follows:

$$Z_{n,L} = Z_{n,p} - Z_{n,q} + \frac{y_{p-q,\rho-\sigma}\, Z_{\rho-\sigma,n}}{y_{p-q,p-q}} \tag{3}$$

and

$$Z_{L,L} = Z_{p,L} - Z_{q,L} + \frac{1+y_{p-q,\rho-\sigma}\, Z_{\rho-\sigma,L}}{y_{p-q,p-q}} \tag{4}$$

The next step in the calculation is the elimination of this new loop row and column by using Kron's formula[8] to modify the original elements of the matrix.

$$Z_{m,n\text{ (after elimination)}} = Z_{m,n\text{ (before elimination)}} - \frac{Z_{mL} Z_{Ln}}{Z_{LL}} \qquad (5)$$

3. If branch p-q is not coupled to the previously assembled subnetwork, then equations 1 - 4 reduce to equations 6 - 9 respectively:

$$Z_{n,q} = Z_{n,p} \qquad (6)$$

$$Z_{q,q} = Z_{p,q} + z_{p-q,p-q} \qquad (7)$$

$$Z_{n,L} = Z_{n,p} - Z_{n,q} \qquad (8)$$

$$Z_{L,L} = Z_{p,L} - Z_{q,L} + z_{p-q,p-q} \qquad (9)$$

4. If terminal p is the reference bus, then:

$$Z_{n,p} = Z_{p,q} = Z_{p,L} = 0$$

ALGORITHM TO COMPUTE Y_{Loop}

A similar (dual) stepwise algorithm for computing the loop-admittance-matrix may be derived in which an added link branch adds a new coordinate while an added tree branch adds new rows and columns which are eliminated. The formulas will be the duals of equations 1 - 4 and 6 - 9 with indices denoting links (loops) instead of buses. This adds some complication since a tree branch may be identified by more than two indices due to the fact that it may be common to more than two loops.

ALGORITHMS TO COMPUTE Y_{LOOP} FROM Z_{BUS} AND VICE VERSA

1. Computing Elements of $Z_{network}$ from Z_{bus}:

$$Z_{p-q,r-s} = Z_{p,r} + Z_{q,s} - Z_{q,r} - Z_{p,s} \qquad (10)$$

2. Computing Elements of $Y_{network}$ from Y_{loop}:

This requires that every network branch be identified by the loops in which it is common considering polarities. If branch p-q is a-b+c and branch r-s is -g+h, then:

$$Y_{p-q,r-s} = Y_{a-b+c,-g+h}$$

$$= -Y_{a,g} + Y_{a,h} + Y_{b,g} - Y_{b,h} - Y_{c,g} + Y_{c,h} \qquad (11)$$

3 Computing Elements of $Y_{network}$ from $Z_{network}$:

$$Y_{p-q,r-s} = y_{p-q,r-s} - y_{p-q,\rho-\sigma}\, Z_{\rho-\sigma,\ \mu-\nu}\, y_{\mu-\nu,r-s} \tag{12}$$

$\rho-\sigma$ and $\mu-\nu$ represent all branches which are mutually coupled to branch p-q including p-q.

4. Computing Y_{loop} from $Y_{network}$:

Grouping the tree branches and the link branches, $Y_{network}$ can be subdivided as follows:

$$Y_{network} = \begin{array}{c|c|c|} & \text{tree branches} & \text{link branches} \\ \hline \text{tree branches} & [Y_1] & [Y_2] \\ \hline \text{link branches} & [Y_3] & [Y_4] \\ \hline \end{array} \tag{13}$$

$$Y_{loop} = Y_{link} = [Y_4] \tag{14}$$

Note that $Y_{network}$ is a redundant matrix where $[Y_1]$, $[Y_2]$ and $[Y_3]$ can be computed from $[Y_4]$ by using equation (11)

5. Computing $Z_{network}$ from $Y_{network}$:

$$Z_{p-q,r-s} = Z_{p-q,r-s} - z_{p-q,\rho-\sigma}\, Y_{\rho-\sigma,\mu-\nu}\, z_{\mu-\nu,r-s} \tag{15}$$

This is the dual of equation (12)

6. Computing Z_{bus} from $Z_{network}$:

Subdividing $Z_{network}$:

$$Z_{network} = \begin{array}{c|c|c|} & \text{tree branches} & \text{link branches} \\ \hline \text{tree branches} & [Z_1] & [Z_2] \\ \hline \text{link branches} & [Z_3] & [Z_4] \\ \hline \end{array} \tag{16}$$

then:

$$Z_{tree\ n.p.} = [Z_1] \tag{17}$$

$$Z_{bus} = K^{-1}\, Z_{tree\ n.p.}\, K_t^{-1} \tag{18}$$

where

K = non-singular transformation matrix which expresses tree node-pair currents in terms of bus currents.

ALGORITHMS FOR MODIFYING Z_{BUZ} AND Y_{LOOP}

If the system change is caused by changes in the impedances of a few branches, then Z_{bus} and Y_{loop} modifications are made as follows:

1. Modifying Elements of Z_{bus}:

$$Z_{m,n(new)} = Z_{m,n(old)} + Z_{m,\mu-\nu} \left[\mathbf{Y}_{\mu-\nu,\rho-\sigma}\right] Z_{\rho-\sigma,n} \tag{19}$$

where:

$$\mathbb{Y}_{\mu-\nu,\rho-\sigma} = \left[\mathbb{U} - \Delta y_{\mu-\nu,\rho-\sigma} Z_{\rho-\sigma,\mu-\nu}\right]^{-1} \Delta y_{\mu-\nu,\rho-\sigma} \tag{20}$$

$\mathbb{U}$ = unit matrix

$$\Delta y_{\mu-\nu,\rho-\sigma} = y_{\mu-\nu,\rho-\sigma(\text{old})} - y_{\mu-\nu,\rho-\sigma(\text{new})} \tag{21}$$

$\mu-\nu$ and $\rho-\sigma$ represent branches whose admittances are affected by the system change.

2. Modifying Elements of Y_{loop}:

$$Y_{a,b(\text{new})} = Y_{a,b(\text{old})} + Y_{a,\mu-\nu}\left[\mathbb{Z}_{\mu-\nu,\rho-\sigma}\right] Y_{\rho-\sigma,b} \tag{22}$$

$$\mathbb{Z}_{\mu-\nu,\rho-\sigma} = \left[\mathbb{U} - \Delta z_{\mu-\nu,\rho-\sigma} Y_{\rho-\sigma,\mu-\nu}\right]^{-1} \Delta z_{\mu-\nu,\rho-\sigma} \tag{23}$$

$$\Delta z_{\mu-\nu,\rho-\sigma} = z_{\mu-\nu,\rho-\sigma(\text{old})} - z_{\mu-\nu,\rho-\sigma(\text{new})} \tag{24}$$

$= \text{change in } z_{\text{branch}}$

$\mu-\nu$ and $\rho-\sigma$ represent branches whose impedances are affected by system change.

Note that for the same system change the branches included in modifying Z_{bus} and Y_{loop} may not be identical due to the inverse relationship between z branch and y branch.

CONCLUSIONS

The network approach provides a unified basis for formulating and solving a wide variety of problems on a digital computer, thereby, in effect, enabling it to be converted into a kind of a general simulation and analysis machine with great versatility and flexibility.

The techniques presented in this paper have been applied to the solution of large power networks under steady state and transient conditions. With these methods and the IBM 7090, it has been possible to simulate the operation and study problems of network serving an eight-state area in a fraction to 20 minutes of computer time [22-24]. The success obtained with these applications is due to the development and application of the algorithms of this paper and continued success can be expected from the extension of these concepts. The principal goal in computer application to system problems should be the development of algorithms specifically designed for digital computation which will provide the most usable results in the most efficient manner.

Besides their advantages in computational and analysis applications, the presented methods could be valuable in synthesis. For example, the expressions developed to modify the solution matrices can be used as good sensitivity measure in system design.

REFERENCES

1. Network Topological Formulation of Structural Analysis, S. J. Fenves and F. H. Branin, Jr., Journal of the Structural Div.. Proceeding of the A. S. C. E., vol. 89, No. ST 4. August 1963, Part I pp. 483-514.
2. Analysis of Discrete Physical Systems, H. E. Koening, Y. Tokad, H. K. Kesavan, to be published.
3. Radiation Analysis by the Network Method, A. K. Oppenheim, Trans. Amer. Sc. Mech. Engrs. 78, 725-735 (1956).
4. An Abstract Mathematical Basis for Network Analysis and its Significance in Physics and Engineering, F. H. Branin, Jr., AIEE. paper S-128, April 1961.
5. Industrial Dynamic, J. W. Forrester, The M. I. T. Press and John Wiley and Sons, Inc., New York, 1961.
6. Who Eats Whom?, F. Harary, General Systems, Yearbook of the Society for General Systems Research, Volume VI, 1961.
7. A Relational Theory of Biological Systems I and II, Robert Rosen, The Bulletin of Mathematical Biophysics, 20, pp. 245-60 and 317-41, 1958.
8. Tensor Analysis of Networks, G. Kron, John Wiley and Sons, New York, N. Y., 1939.
9. Linear Graphs and Electrical Networks, S. Seshu and M. B. Reed, Addison Wesley, Reading, Mass., 1961.
10. Matrix Analysis of Electric Networks, P. Le Carbeiller, John Wiley and Sons, Inc., New York, N. Y. (1950).
11. Digital Calculation of Network Inverse and Mesh Transformation Matrices, N. Sato, AIEE Trans.
12. Application of Topology to Network Analysis (Ph. D. Thesis), G. E. Ching, Purdue University, Lafayette, Indiana, August 1960.
13. Diakoptics, the Piecewise Solution of Large-Scale Systems, G. Kron, The Electrical Journal, (London), a monthly series beginning June 1957 to February 1959. Recently published in a bookform by McDonald, London.
14. A Set of Principles to interconnect the solutions of physical systems, G. Kron, J. Appl. Phys. 24 (1953) pp. 965-980.
15. A method to solve very large Physical Systems in Easy Stages, G. Kron, Proceedings, IRE Vol. 42, 1954, pp. 680-686.
16. An application of algebraic topology: Kron's method of tearing, J. Paul Roth, Quart. Appl. Math. 17 (1959), pp. 1-24.
17. A Critical Study of Kron's Method of Tearing, R. R. Sabroff and J. J. Higgins, The Matrix and Tensor Quarterly, vol. 7, No. 4, June 57 to Vol. 9, No. 1 September 1958.
18. A Discussion of Mathematical Techniques Used in Kron's Method of Tearing, B. Kent Harrison, Journal of the Society of Industrial and Applied Mathematics, Vol. 11, No. 2, pp. 258-280, June 1963.
19. Recursive Methods, Dynamic Programming and Invariant Imbedding, Arthm Wouk, SIAM Review, vol. 4, No. 4, October, 1962.

20. Principles of Numerical Analysis, A. S. Householder, McGraw-Hill, New York, 1953, pp. 79.

21. Digital Calculation of L-G Short Circuits by Matrix Method, A. H. El-Abiad, AIEE Transactions, vol. 709, pt. III, 1960, pp. 323-32.

22. Calculation of Short Circuits Using the High-Speed Digital Computer, A. H. El-Abiad, Ruth Guidene and G. W. Stagg, AIEE Transactions, Power Apparatus and Systems, December 1961, pp. 702-708.

23. Automatic Evaluation of Power System Performance - Effect of Line and Transformer Outages," A. H. El-Abiad, G. W. Stagg, AIEE Transactions, Power Apparatus and Systems, February, 1963, pp. 712-716.

24. The Load Flow Problem, Its Formulation and Solution, Proceedings of the Power Systems Computation Conference, Queen Mary College, London, September 16-20, 1963.

APPENDIX

NUMERICAL EXAMPLE

Given for the sample system of Figure 3.

Step	p-q	$z_{p-q,p-q}$	r-s	$z_{p-q,r-s}$
1	1-2(1)	.6		
2	1-2(2)	.4	1-2(1)	.2
3	1-3	.5	1-2(1)	.1
4	3-4	.5		
5	2-4	.2		

To compute Z_{bus}

1. Start with bus 1. Add branch 1-2 (1) $\begin{array}{c|c} & 2 \\ \hline 2 & .6 \end{array}$
2. Add branch 1-2 (2). This closes a loop. Use equations 3 and 4. Due to mutual terms, make a $z_{p-q,r-s}$ matrix and invert for y terms.

$$z = \begin{array}{c|cc} & 1\text{-}2\ (1) & 1\text{-}2\ (2) \\ \hline 1\text{-}2\ (1) & .6 & .2 \\ 1\text{-}2\ (2) & .2 & .4 \end{array}$$

$$y = \begin{array}{c|cc} & 1\text{-}2\ (1) & 1\text{-}2\ (2) \\ \hline 1\text{-}2\ (1) & 2 & -1 \\ 1\text{-}2\ (2) & -1 & 3 \end{array}$$

$p = 1,\ q = 2,\ n = 2$

$\rho = 1,\ \sigma = 2$

$$Z_{n,L} = Z_{n,p} - Z_{n,q} + \frac{y_{p-q,\rho-\sigma}\ Z_{\rho-\sigma,n}}{y_{p-q,p-q}};\quad Z_{\rho-\sigma,n} = Z_{\rho,n} - Z_{\sigma,n}$$

$$Z_{2,L} = Z_{2,1} - Z_{2,2} + \frac{y_{1-2\ (2),1-2\ (1)}\ (Z_{1,2} - Z_{2,2})}{y_{1-2\ (2),\ 1-2\ (2)}}$$

$$= 0 - .6 + \frac{-1\ \ (0 - .6)}{3} = -.4$$

$$Z_{L,L} = Z_{p,L} - Z_{q,L} + \frac{1 + y_{p-q,\ \rho-\sigma}\ Z_{\rho-\sigma,\ L}}{y_{p-q,\ p-q}};\quad Z_{\rho-\sigma,L} = Z_{\rho,L} - Z_{\sigma,L}$$

$$Z_{L,L} = Z_{1,L} - Z_{2,L} + \frac{1 + y_{1-2\ (2),\ 1-2\ (1)}\ (Z_{1,L} - Z_{2,L})}{y_{1-2\ (1),\ 1-2\ (2)}}$$

$$= 0 - (-.4) + \frac{1 + (-1)\ [0 - (-.4)]}{3} = .6 \qquad \begin{array}{c|c|c} & 2 & L \\ \hline 2 & .6 & -.4 \\ \hline L & -.4 & .6 \end{array}$$

Eliminate L: $\begin{array}{c|c|} & 2 \\ \hline 2 & .333 \\ \hline \end{array}$

3. Add Branch 1-3. This adds Bus 3. Use equations 1 and 2 again. z and y matrices are needed.

$$z = \begin{array}{c|c|c|c|} & 1\text{-}2\ (1) & 1\text{-}2\ (2) & 1\text{-}3 \\ \hline 1\text{-}2\ (1) & .6 & .2 & .1 \\ \hline 1\text{-}2\ (2) & .2 & .4 & \\ \hline 1\text{-}3 & .1 & & .5 \\ \hline \end{array}$$

$$y = \begin{array}{c|c|c|c|} & 1\text{-}2\ (1) & 1\text{-}2\ (2) & 1\text{-}3 \\ \hline 1\text{-}2\ (1) & 2.08 & -1.04 & -.417 \\ \hline 1\text{-}2\ (2) & -1.04 & 3.06 & .208 \\ \hline 1\text{-}3 & -.417 & .208 & 2.08 \\ \hline \end{array}$$

$p = 1,\ q = 3,\ n = 2,\ \rho = 1,\ \sigma = 2$

$$Z_{n,p} = Z_{n,p} + \frac{y_{p-q,\ \rho-\sigma}\ Z_{\rho-\sigma,n}}{y_{p-q,p-q}};\quad Z_{\rho-\sigma,n} = Z_{\rho,n} - Z_{\sigma,n}$$

$$Z_{2,3} = Z_{2,1} + \frac{\begin{bmatrix} y_{1-3,\ 1-2\ (1)} & y_{1-3,\ 1-2\ (2)} \end{bmatrix} \begin{bmatrix} Z_{1,2} - Z_{2,2} \\ Z_{1,2} - Z_{2,2} \end{bmatrix}}{y_{1-3,\ 1-3}}$$

$$= 0 + \frac{\begin{bmatrix} -.417 & .208 \end{bmatrix} \begin{bmatrix} 0 - -.333 \\ 0 - -.333 \end{bmatrix}}{2.08}$$

$$Z_{q,q} = Z_{p,q} + \frac{1 + y_{p-q,\ \rho-\sigma}\ Z_{\rho-\sigma,q}}{y_{p-q,p-q}};\quad Z_{\rho-\sigma,q} = Z_{\rho,q} - Z_{\sigma,q}$$

$$Z_{3,3} = Z_{1,3} + \frac{1 + \begin{bmatrix} y_{1-3,\ 1-2\ (1)} & y_{1-3,\ 1-2\ (2)} \end{bmatrix} \begin{bmatrix} Z_{1,3} - Z_{2,3} \\ Z_{1,3} - Z_{2,3} \end{bmatrix}}{y_{1-3,\ 1-3}}$$

$$= 0 + \frac{1 + \begin{bmatrix} -.417 & .208 \end{bmatrix} \begin{bmatrix} 0 - .033 \\ 0 - .033 \end{bmatrix}}{2.08} = .483$$

$$\begin{array}{c|c|c|}
 & 2 & 3 \\
\hline
2 & .333 & .033 \\
\hline
3 & .033 & .483 \\
\hline
\end{array}$$

4. Add Branch 3 - 4. This adds bus 4. Use equations (6) and (7)

$$\begin{array}{c|c|c|c|}
 & 2 & 3 & 4 \\
\hline
2 & .333 & .033 & .033 \\
\hline
3 & .033 & .483 & .483 \\
\hline
4 & .033 & .483 & .983 \\
\hline
\end{array}$$

5. Add Branch 2 - 4. This closes a loop. Use equations (8) and (9).

$$\begin{array}{c|cccc|}
 & 2 & 3 & 4 & L \\
\hline
2 & .333 & .033 & .033 & .3 \\
\hline
3 & .033 & .483 & .483 & -.45 \\
\hline
4 & .033 & .483 & .983 & -.95 \\
\hline
L & .3 & -.45 & -.95 & 1.45 \\
\hline
\end{array}$$

Use equation (5)
to eliminate L

$$Z_{bus} = \begin{array}{c|ccc|}
 & 2 & 3 & 4 \\
\hline
2 & .272 & .126 & .230 \\
\hline
3 & .126 & .343 & .188 \\
\hline
4 & .230 & .188 & .360 \\
\hline
\end{array}$$

To Compute Y_{loop} From Z_{bus}

Coding branches by loops:

Branch	Coding by bus	Coding by loop
1	1-2 (1)	-a+b
2	1-2 (2)	a
3	1-3	-b (1)
4	3-4	-b (2)
5	2-4	b

Using Equation (12):

$$Y_{a,a} = Y_{1-2\ (2).\ 1-2\ (2)}$$

$$= y_{1-2\ (2),\ 1-2\ (2)} - y_{1-2\ (2),\ \rho-\sigma}\, Z_{\rho-\sigma,\mu-\nu}\, y_{\mu-\nu,\ 1-2\ (2)}$$

where

$$y_{1-2\ (2),\ 1-2\ (2)} = 3.02$$

$$y_{1-2\ (2),\ \rho-\sigma} = \begin{array}{c|c|c|c|}
 & 1-2\ (1) & 1-2\ (2) & 1-3 \\
\hline
1-2\ (2) & -1.04 & 3.03 & .208 \\
\hline
\end{array}$$

$$Z_{\rho-\sigma,\ \mu-\nu} = \begin{array}{c|c|c|c|} & 1\text{-}2\ (1) & 1\text{-}2\ (2) & 1\text{-}3 \\ \hline 1\text{-}2\ (1) & .271 & .271 & .126 \\ \hline 1\text{-}2\ (2) & .271 & .271 & .126 \\ \hline 1\text{-}3 & .126 & .126 & .343 \\ \hline \end{array}$$

$$y_{\mu-\nu,\ 1-2\ (2)} = \begin{array}{c|c|} & 1\text{-}2\ (2) \\ \hline 1\text{-}2\ (1) & -1.04 \\ \hline 1\text{-}2\ (2) & 3.02 \\ \hline 1\text{-}3 & .208 \\ \hline \end{array}$$

which after proper matrix multiplication results in

$$Y_{a,a} = 1.84$$

$$Y_{a,b} = Y_{1-2\ (2),\ 2-4}$$

$$= y_{1-2\ (2),\ 2-4} - y_{1-2\ (2),\ \rho-\sigma}\ Z_{\rho-\sigma,\mu-\nu}\ y_{\mu-\nu,2-4}$$

$$Y_{a,b} = 0 - 1\text{-}2\ (2)\ \begin{array}{|c|c|c|} 1\text{-}2\ (1) & 1\text{-}2\ (2) & 1\text{-}3 \\ \hline -1.04 & 3.02 & .208 \\ \hline \end{array}\ \begin{array}{c|c|} & 2\text{-}4 \\ \hline 1\text{-}2(1) & .230\ -.271 \\ 1\text{-}2(2) & .230\ -.271 \\ 1\text{-}3 & .188\ -.126 \\ \hline \end{array}\ 2\text{-}4\ \begin{array}{|c|} 2\text{-}4 \\ \hline 5.0 \\ \hline \end{array}$$

$$= -.345$$

$$Y_{b,b} = Y_{2-4,\ 2-4}$$

$$= y_{2-4,\ 2-4} - y_{2-4,\ 2-4}\ Z_{2-4,\ 2-4}\ y_{2-4,\ 2-4}$$

$$= 5 - 5\ (.272 + .360 - .46)\ 5$$

$$= .7$$

$$\therefore\ Y_{loop} = \begin{array}{c|c|c|} & a & b \\ \hline a & 1.84 & -.345 \\ \hline b & -.345 & 0.7 \\ \hline \end{array}$$

Matrix Modification

Find the new Z_{bus} and Y_{loop} due to the following changes in the sample system:

$z_{1-2\ (1),\ 1-2\ (1)}$ change from 0.6 to 0.4

and

$z_{1-2\ (1),\ 1-2\ (2)}$ change from 0.2 to 0

$$z_{\text{branch (old)}} = \begin{array}{c|ccccc} & \text{1-2 (1)} & \text{1-2(2)} & \text{1-3} & \text{3-4} & \text{2-4} \\ \hline \text{1-2 (1)} & .6 & .2 & .1 & & \\ \text{1-2 (2)} & .2 & .4 & & & \\ \text{1-3} & .1 & & .5 & & \\ \text{3-4} & & & & .5 & \\ \text{2-4} & & & & & .2 \end{array}$$

$$y_{\text{branch (old)}} = \begin{bmatrix} 2.08 & -1.04 & -.417 & & \\ -1.04 & 3.02 & .208 & & \\ -.417 & .208 & 2.08 & & \\ & & & 2 & \\ & & & & 5 \end{bmatrix}$$

$$z_{\text{branch (new)}} = \begin{bmatrix} .4 & & .1 & & \\ & .4 & & & \\ .1 & & .5 & & \\ & & & .5 & \\ & & & & .2 \end{bmatrix}$$

$$y_{\text{branch (new)}} = \begin{bmatrix} 2.63 & & -.526 & & \\ & 2.5 & & & \\ -.526 & & 2.11 & & \\ & & & 2 & \\ & & & & 5 \end{bmatrix}$$

$$\Delta z_{\mu-\nu,\rho-\sigma} = z_{\text{branch (old)}} - z_{\text{branch (new)}}$$

$$= \begin{array}{c|cc} & \text{1-2 (1)} & \text{1-2 (2)} \\ \hline \text{1-2 (1)} & .2 & .2 \\ \text{1-2 (2)} & .2 & 0 \end{array}$$

$$\Delta y_{u-\nu,\rho-\sigma} = y_{\text{branch (old)}} - y_{\text{branch (new)}}$$

$$= \begin{array}{c|ccc} & \text{1-2 (1)} & \text{1-2 (2)} & \text{1-3} \\ \hline \text{1-2 (1)} & -.55 & -1.04 & .11 \\ \text{1-2 (2)} & -1.04 & .52 & .208 \\ \text{1-3} & .11 & .208 & -.03 \end{array}$$

Use equations 19 - 21 to find $Z_{\text{bus (new)}}$

$$\mathbb{Y}_{\mu-\nu,\rho-\sigma} = [\mathbb{U} - \Delta y_{\mu-\nu,\rho-\sigma}\, Z_{\mu-\nu,\rho-\sigma}]^{-1}\, \Delta y_{\mu-\nu,\rho-\sigma}$$

$$\left[\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} - \begin{bmatrix} -.55 & -1.04 & .11 \\ -1.04 & .52 & .208 \\ .11 & .208 & -.03 \end{bmatrix} \begin{bmatrix} .272 & .272 & .126 \\ .272 & .272 & .126 \\ .126 & .126 & .343 \end{bmatrix} \right]^{-1} \begin{bmatrix} -.55 & -1.04 & .11 \\ -1.04 & .52 & .208 \\ .11 & .208 & -.03 \end{bmatrix}$$

$$= \begin{array}{c|ccc} & \text{1-2 (1)} & \text{1-2 (2)} & \text{1-3} \\ \hline \text{1-2 (1)} & -.1193 & -.926 & .0241 \\ \text{1-2 (2)} & -.926 & .569 & .1841 \\ \text{1-3} & .0241 & .1841 & -.0033 \end{array}$$

$$Z_{m,n(new)} = Z_{m,n(old)} + Z_{m,\mu-\nu}\,\mathbb{Y}_{\mu-\nu,\rho-\sigma}\,Z_{\rho-\sigma,m}$$

$$Z_{22(new)} = .272 + 2 \begin{array}{ccc} \text{1-2(1)} & \text{1-2(2)} & \text{1-3} \\ \hline -.272 & -.272 & -.126 \end{array} \begin{bmatrix} -.1193 & -.926 & .0241 \\ -.926 & .569 & .1841 \\ .0241 & .1841 & -.0033 \end{bmatrix} \begin{bmatrix} -.272 \\ -.272 \\ -.126 \end{bmatrix}$$

$$= .185$$

Similarly $Z_{23(new)} = .101$

$Z_{24(new)} = .156$

$Z_{33(new)} = .338$

$Z_{34(new)} = .169$

$Z_{44(new)} = .304$

or

$$Z_{bus(new)} = \begin{array}{c|ccc} & 2 & 3 & 4 \\ \hline 2 & .185 & .101 & .156 \\ 3 & .101 & .338 & .169 \\ 4 & .156 & .169 & .304 \end{array}$$

Use Equations 22-24 to Find $Y_{loop\ (new)}$:

$$\mathbb{Z}_{\mu- ,\rho-\sigma} = [\mathbb{U} - \Delta z_{\mu-\nu,\rho-\sigma}\, Y_{\mu-\nu,\rho-\sigma}]^{-1}\, \Delta z_{\mu-\nu,\rho-\sigma}$$

$$= \left[\begin{bmatrix} 1 & \\ & 1 \end{bmatrix} - \begin{bmatrix} .2 & .2 \\ .2 & 0 \end{bmatrix} \begin{bmatrix} 1.839 & -1.495 \\ -1.495 & 1.83 \end{bmatrix} \right]^{-1} \begin{bmatrix} .2 & .2 \\ .2 & 0 \end{bmatrix} = \begin{array}{c|cc} & \text{1-2(1)} & \text{1-2(2)} \\ \hline \text{1-2(1)} & .230 & .219 \\ \text{1-2(2)} & .219 & .062 \end{array}$$

$$Y_{ab(new)} = Y_{ab(old)} + Y_{a,\mu-\nu}\,\mathbb{Z}_{\mu-\nu,\rho-\sigma}\,Y_{\rho-\sigma,b}$$

$$= -.345 + a \begin{array}{cc} \text{1-2(1)} & \text{1-2(2)} \\ \hline -1.495 & 1.84 \end{array} \begin{bmatrix} .230 & .219 \\ .219 & .062 \end{bmatrix} \begin{bmatrix} .344 \\ .345 \end{bmatrix}$$

$$= .293$$

Similarly:

$Y_{aa} = 1.358$

$Y_{bb} = .776$

or

$$Y_{loop(new)} = \begin{array}{c|cc} & a & b \\ \hline a & 1.358 & .293 \\ b & .293 & .776 \end{array}$$

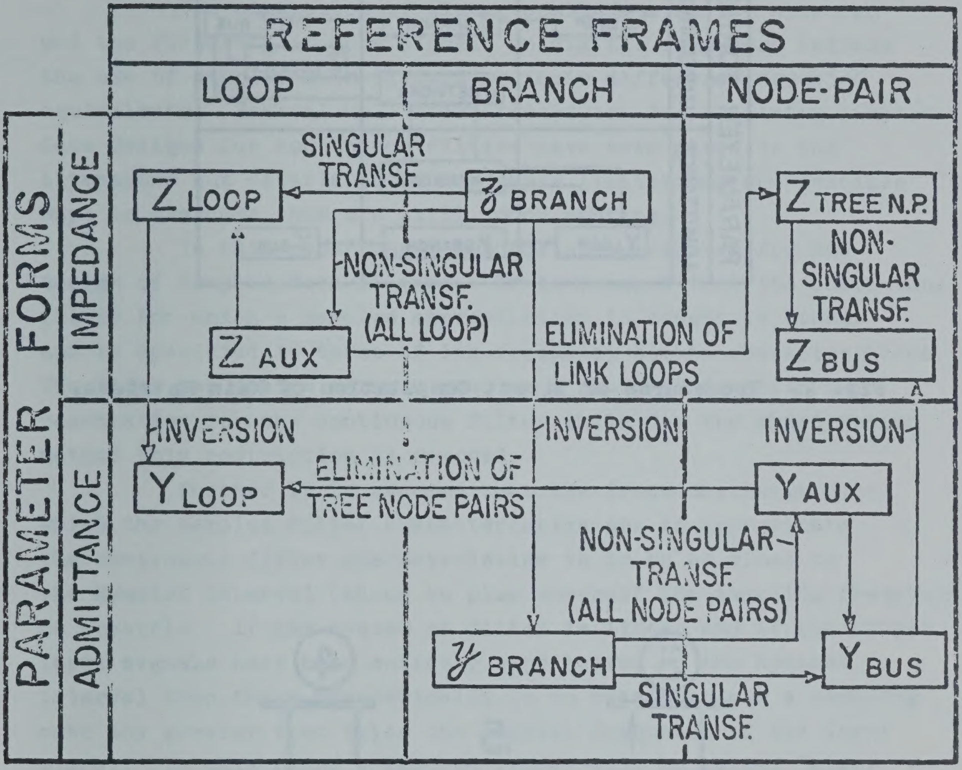

REFERENCE FRAMES
LOOP
BRANCH
NODE-PAIR
PARAMETER FORMS
IMPEDANCE
ADMITTANCE
SINGULAR TRANSF.
Z LOOP
z BRANCH
Z TREE N.P.
NON-SINGULAR TRANSF. (ALL LOOP)
NON-SINGULAR TRANSF.
Z AUX
ELIMINATION OF LINK LOOPS
Z BUS
INVERSION
INVERSION
INVERSION
ELIMINATION OF TREE NODE PAIRS
Y LOOP
Y AUX
NON-SINGULAR TRANSF. (ALL NODE PAIRS)
y BRANCH
Y BUS
SINGULAR TRANSF.

Fig. 1. Classification of Network Main Matrices by Parameter Forms and Reference Frames. Flow Lines Indicate Procedures of Derivation by Classical Methods.

Zaux: Represents the Network With Added Infinite Impedance Ficticious Branches Across Tree Branches Giving An All-Loop Auxiliary Network.

Yaux: Represents the Network With Added Zero Impedance Ficticious Branches In Series With Link Branches Giving An All-Node Pair Auxiliary Network.

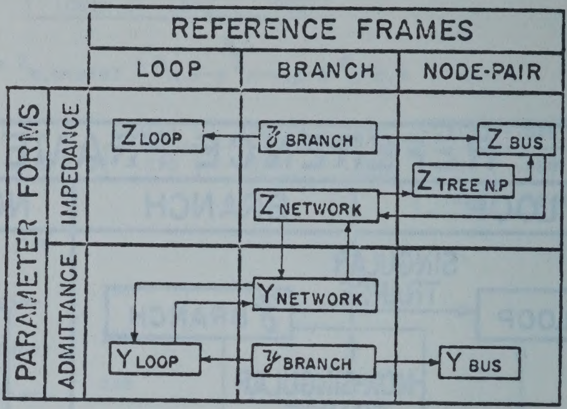

REFERENCE FRAMES
LOOP
BRANCH
NODE-PAIR
PARAMETER FORMS
IMPEDANCE
ADMITTANCE
Z LOOP
z BRANCH
Z BUS
Z TREE N.P
Z NETWORK
Y NETWORK
Y LOOP
y BRANCH
Y BUS

Fig. 2. Procedures of Direct Computation of Main Matrices.

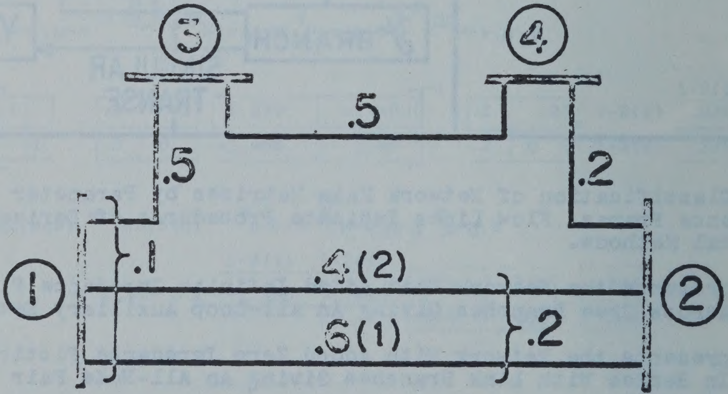

③
④
.5
.5
.2
.1
.4(2)
①
②
.6(1)
.2

Fig. 3. Sample System.

DESIGN METHODS FOR SAMPLED DATA FILTERS

J. F. KAISER
Bell Telephone Laboratories
Murray Hill, New Jersey

Introduction

The simulation of linear continuous filter networks and the filtering of data signals on digital computers require the use of sampled data filters or their difference equation equivalents. Several well known procedures for obtaining sampled data designs for continuous filters have been given in the literature but details outlining their limitations and possible ways to overcome them are still found wanting.

In this paper three methods are presented for the design of sampled data filters. It is assumed that the continuous filter for which a sampled approximation is sought is linear and is specified in terms of its frequency domain characteristics. The first two design methods require a rational function representation for the continuous filter while for the third design method this restriction is removed.

Further it is assumed that the frequency range over which the sampled filter characteristics are to approximate the continuous filter characteristics is to be as close to the Nyquist interval (minus to plus one-half the sampling frequency) as possible. If the system or filter is linear and if the filter input signals have been suitably bandlimited to the Nyquist interval then there theoretically is no reason to use a sampling rate any greater than twice the highest frequency of the input signal. In this paper a sampled filter will be termed "wideband" if the frequency range of useful approximation approaches the Nyquist limit. Thus the sampling frequency of a "wideband" filter will usually be from two to two and one-half times the highest data or signal frequency of interest. This is in somewhat sharp contrast to the sampled filters used in control system design where the sampling rate is usually chosen to lie between five and twelve or more times the effective control system bandwidth so as not to introduce excessive amounts of delay in the control loop.

The use of a minimal sampling rate will in general result in a minimum total amount of computation to be performed in any simulation or data processing operation unless, of course, the use of this sampling rate requires an inordinately complex sampled filter design. The relationship between sampled filter complexity as measured by the number of filter coefficients and the effective bandwidth of useful approximation is indicated for each design method.

The first design method employs the standard z-transformation with modifications where necessary. The use of the bilinear z-transformation with suitable scale modification constitutes the second method. A procedure utilizing a Fourier series expansion of the filter magnitude characteristics with truncation errors minimized by using spectral windows is the method advanced for the design of nonrecursive sampled data filters.

The Standard Z-Transform

The nature of the approximation of the standard z-transform[1] can be seen from the defining equations for the transform. Let H(s) be the transfer function (the Laplace transform of its impulse response) of a linear continuous filter. Then the standard z-transform of this transfer characteristic is defined by

$$H^*(s) = \sum_{n=-\infty}^{\infty} H(s+jn\omega_s) \qquad (1)$$

or equivalently by

$$H^*(z) = T \sum_{n=0}^{\infty} h(nT)z^{-n} \qquad (2)$$

where:

$s = \sigma + j\omega$

$\omega_s = \frac{2\pi}{T}$ = the radian sampling frequency

$z^{-1} = e^{-sT}$ the unit delay operator

$h(t)$ = impulse response of the continuous filter

The asymptotic behavior for large s is assumed to be

---

[1] See: Wilts, C. H., "Principles of Feedback Control," Addison-Wesley, Reading, Mass. 1960

$$\lim_{s \to j\infty} |H(s)| = \lim_{s \to j\infty} |1/(s/\omega_c)^n|, \quad n > 0 \qquad (3)$$

Thus, from (2), the z-transform is so defined that the impulse response of the sampled filter is identical to the sampled impulse response of the continuous filter. Further, H*(s) is periodic in ω of period ω_s. From (1) it follows that in the baseband $(-\omega_s/2 \leq \omega \leq \omega_s/2)$ the frequency response characteristics of the obtained sampled filter, H*(s), differ from those of the continuous filter, H(s), the difference being the amount added or "folded" in through terms of the form $H(s+jn\omega_s)$, $n \neq 0$. If H(s) is bandlimited to the baseband, i.e., $|H(s)| = 0$ for $|\omega| > \omega_s/2$, then there is no folding error and the frequency response of the sampled-data filter is identical to that of the continuous filter. Unfortunately, when H(s) is a rational function of s, it is not bandlimited and therefore $H(s) \neq H^*(s)$ in the baseband.

The magnitude of the errors resulting from the folding is directly related to the high frequency asymptotic behavior of H(s) as defined in Eq. 3. If n is large and $\omega_c \ll \omega_s/2$, then the folding errors will be small and the standard z-transform generally will yield a satisfactory sampled-data filter design. However, in wideband simulations, ω_c is usually an appreciable fraction of $\omega_s/2$. Furthermore, many continuous filter designs, such as elliptic filters or Chebyshev Type II filters, yield transfer functions in which n is no greater than 1. These two conditions, namely $\omega_c \approx \omega_s/2$ and $n = 1$, can create large folding errors in the frequency response characteristics of the filter obtained by using the standard z-transform and thus render an unusable result.

To reduce the possibility of error due to folding the H(s) can be modified by adding in cascade a wideband low-pass filter G(s), having an 'n' sufficiently large. This "guard" filter, G(s), is usually chosen to have flat magnitude and linear phase characteristics in the frequency range where it is desired to have the H*(s) faithfully approximate H(s). This may be obtained by using one of the standard low-pass all-pole filter forms such as the Butterworth or Chebyshev Type I and if necessary followed by an all-pass filter for phase equalization.

The procedure is to form $H_m(s)$ as

$$H_m(s) = G(s)\,H(s) \qquad (4)$$

and then to apply the standard z-transformation to this $H_m(s)$.

With H(s) the quotient of a numerator polynomial N(s) of order n_N and a denominator polynomial D(s) of order n_D where $n_D > n_N$, application of the standard z-transformation requires first determining the zeros of D(s), then obtaining a partial fraction expansion of H(s) in these poles, and finally z-transforming each of the individual terms. From the following transform pair all the necessary pairs can be derived.

$$\frac{1}{s+a} \Longrightarrow \frac{T}{1-e^{-aT}z^{-1}} \qquad (5)$$

Thus it is seen that H*(z) will be rational in z^{-1} and will have at most $2n_D$ coefficients. If in addition a "guard" filter of order n_G is used then the obtained $H^*_m(z)$ will be more complex now having at most $2(n_D+n_G)$ coefficients.

Another limiting feature of the standard z-transformation is the fact that the moments of h(t) do not remain invariant under z-transformation, i.e., in general.

$$\sum_{\ell=0}^{\infty} (\ell T)^n\, h(\ell T) \neq \int_0^{\infty} t^n h(t)\,dt \qquad (6)$$

However, the numerator coefficients of H*(z) may usually be varied slightly so as to satisfy an equality for the first few moments without seriously affecting the overall frequency characteristics of H*(z).

The Bilinear Z-Transformation

To circumvent the "folding" problem of the standard z-transform, a transformation can be employed which will map the entire complex s plane into the horizontal strip in the s_1 plane bounded by the lines $s_1 = -j\omega_s/2$ and $s_1 = +j\omega_s/2$. Since a permissible H*(z) must also be periodic in ω of period ω_s this transformation must also cause H(s) to be mapped identically in each of the other horizontal strips bounded by the lines $s_1 = j(n-\frac{1}{2})\,\omega_s$ and $s_1 = j(n+\frac{1}{2})\,\omega_s$ where n is an integer.

A transformation having these properties is the bilinear z-transform or z-form which is defined as

$$s = \frac{2}{T} \tanh \left(\frac{s_1 T}{2}\right) \qquad (7)$$

which becomes upon substituting

$$z^{-1} = e^{-s_1 T}$$

$$s = \frac{2}{T} \frac{(1-z^{-1})}{(1+z^{-1})} \qquad (8)$$

Thus in terms of the z^{-1} plane this algebraic transformation (8) has the property of uniquely mapping the entire left half of the s plane into the exterior of the unit circle in the z^{-1} plane. Folding errors are eliminated since no folding occurs.

The z-form is applied simply by making the substitution indicated by (8) in the transfer characteristic H(s). Hence

$$H^*(z) = H(s) \Bigg|_{s = \frac{2}{T} \frac{(1-z^{-1})}{(1+z^{-1})}} \qquad (9)$$

The price paid for this feature is the nonlinear warping imparted to the frequency scale as can be seen by setting $s_1 = j\omega_1$ and $s = j\omega$ in (7) to yield

$$\frac{\omega T}{2} = \tan \left(\frac{\omega_1 T}{2}\right) \qquad (10)$$

The deviation from linearity of this relation is shown in Figure 1.

Because of this frequency warping aspect, the z-form is most useful in obtaining sampled-data filter approximations for continuous filters whose magnitude characteristics can be divided along the frequency scale into successive pass and stop bands where the gain or loss is essentially constant in each band. This type of frequency behavior is typical of many low-pass, bandpass and bandstop filters. Compensation can be made for the effect of warping by prewarping the continuous filter design in the opposite way such that upon applying the z-form the critical frequencies will be shifted back to the desired values.

On the other hand if the magnitude frequency characteristics are not essentially piecewise constant over the major portion of the Nyquist frequency interval, such as for a wideband differentiating filter, the frequency warping inherent in the z-form method can render an unsatisfactory filter design.

Thus the z-form has two primary advantages. First, the magnitude frequency characteristics of the continuous filter are carried directly over to the obtained sampled-data filter except for a warping of the frequency scale, i.e., continuous filters with equiripple magnitude characteristics transform into equiripple sampled filters with only the position of the maxima and minima being shifted in frequency a calcuable and compensatable amount. Secondly, application of the z-form is algebraic in form and can be applied equally well to the rational transfer characteristics of the continuous filter in either polynomial or factored form.

With respect to the moments of h(t) it can be shown that only the first two non-zero moments remain invariant under application of the bilinear z-transformation. As in the case of the standard z-transform, agreement can be forced for higher order moments if slight perturbations are made in the numerator coefficients of $H^*(z)$.

In comparing the z-form method with the standard z-transform method it is seen that both methods yield sampled-data transfer functions which are rational in z^{-1} and of the same degree namely that of the degree of the continuous filter. Further when the ratio of the moduli of the poles of the continuous filter to the sampling frequency approaches zero the two methods give identical results as one would expect. For obtaining "wideband" designs by use of the standard z-transformation guard filters must usually be used thus raising the order of the sampled filter. For obtaining "wideband" designs using the z-form the frequency warping must always be compensated for; although no increase in filter order is necessary this method is restricted to filters with essentially piecewise constant magnitude characteristics.

In those situations where a rational approximation is not known for the desired magnitude characteristics of

H(s) or where the magnitude specifications cover the entire Nyquist sampling interval ($|\omega| \leq \omega_s/2$) and where neither the standard z or bilinear z-transformations can be readily applied, the design procedure described in the next section may be used to advantage.

A Nonrecursive Filter Method

The method consists first in expanding the magnitude frequency characteristics desired of the continuous filter H(s) for the band $|\omega| < \omega_s/2$ in a Fourier series about zero frequency. The error resulting from using only a finite number of terms in this expansion is then reduced by applying a suitable weighting function to the expansion coefficients. Finally the individual cosine or sine terms in the expansion are written in terms of the unit delay operator, z^{-1}, by using the exponential forms for the trigonometric functions. The result is then a nonrecursive sampled-data filter.

Since the Fourier expansion is made about zero frequency, the choice of a sine or cosine series is made on the basis of the behavior of H(s) at very small frequencies. If $H(s) \approx Ks^m$ for small s then a cosine series is chosen when m is even and a sine series when m is odd. Thus

$$H(\omega) = \sum_{n=0}^{\infty} a_n \cos n\omega T, \text{ m even} \tag{11a}$$

or

$$H(\omega) = \sum_{n=1}^{\infty} b_n \sin n\omega T, \text{ m odd} \tag{11b}$$

but by definition

$$z^{-1} = e^{-j\omega T} \tag{12}$$

hence

$$H^*(z) = a_o + \tfrac{1}{2} \sum_{n=1}^{\infty} a_n (z^n + z^{-n}), \quad \text{m even} \tag{13a}$$

and

$$H^*(z) = \tfrac{1}{2} \sum_{n=1}^{\infty} b_n (z^n - z^{-n}), \quad \text{m odd} \tag{13b}$$

If the convergence of the Fourier series is not sufficiently rapid enough to make errors resulting from truncating the infinite series of (13) small then in order to yield a satisfactory approximation for (13) with a finite number of terms the coefficients a_n or b_n must be modified. This can be conveniently accomplished by multiplying the time response h(nT) by a time limited even function w(t), i.e., $w(t) = w(-t)$ and $w(t) = 0$ for $|t| > \tau$.

Multiplication in the time domain corresponds to convolution in the frequency domain. Thus, if the weighting function, w(t), is chosen such that the frequency content of its transform is concentrated primarily in the central lobe, i.e., $|\omega| \leq \omega_o$ then the effect of this convolution on the frequency characteristics of H*(z) will be to smooth out the sharp transititions or slope discontinuities in H*(z). This procedure is sketched in Fig. 2 where the desired filter is a wideband differentiator.

The design equations thus become

$$H_1{}^*(z) = a_o\, w(o) + \tfrac{1}{2} \sum_{n=1}^{N} [a_n w(nT)][z^n + z^{-n}], \quad m \text{ even} \qquad (14a)$$

or

$$H_1{}^*(z) = \tfrac{1}{2} \sum_{n=1}^{N} [b_n w(nT)][z^n - z^{-n}], \quad m \text{ odd} \qquad (14b)$$

where N is the greatest integer in (τ/T).

An especially simple weighting function which may be used is Hamming's[2] window function defined by

$$w_h(t) = \begin{cases} 0.54 + 0.46 \cos(\pi t/\tau), & |t| < \tau \\ 0, & |t| > \tau \end{cases} \qquad (15)$$

For this function 99.96% of its energy lies in the band $|\omega| \leq 2\pi/\tau$ with the peak amplitude of the side lobes of $W_h(j\omega)$ being less than 1% of the peak.

A family of weighting functions with nearly optimum characteristics is given by the Fourier cosine transform pair

---

[2] Blackman, R. B. and J. W. Tukey, "The Measurement of Power Spectra," Dover Publications, Inc., New York, 1958, pp. 95-99.

$$w(t) = \begin{cases} \dfrac{I_0\,[\omega_a\sqrt{\tau^2-t^2}\,]}{I_0(\omega_a\tau)} & |t| < \tau \\ 0 & |t| > \tau \end{cases} \qquad (16a)$$

$$W(j\omega) = \frac{2}{I_0(\omega_a\tau)}\ \frac{\sin\tau\sqrt{\omega^2-\omega_a^2}}{\sqrt{\omega^2-\omega_a^2}} \qquad (16b)$$

where I_0 is the modified Bessel function of the first kind and order zero. By varying the product $\omega_a\tau$ the energy in the central lobe can be changed. The usual range on values of $\omega_a\tau$ is $4 < \omega_a\tau < 9$ corresponding to a range of side lobe peak heights of 3.1% down to 0.047%. Fig. 3 shows w(t) and $W(j\omega)$ for $\omega_a\tau = 6.0$. These functions can be shown to approximate closely the prolate spheroidal wave functions of order zero and whose bandlimiting properties[3] are well known.

By viewing this scheme for increasing the convergence of Fourier series at points removed from discontinuities, the relation between the number of terms required in the sampled filter $H^*(z)$ and the usable frequency range of the filter becomes clear. Referring to Fig. 2 again, the halving of ω_0 and the concomitant doubling of τ yield a filter design having twice as many terms as before.

This nonrecursive filter design procedure is best applied to filters whose magnitude characteristics are fairly well behaved especially at very small frequencies and at frequencies approaching the Nyquist limit. For "wideband" filter designs this procedure is superior to the standard numerical analysis approach which usually consists of a Taylor series expansion about zero frequency of the desired magnitude characteristics the result being expressed in differences of increasing order.

Examples

To illustrate the range of the design methods put forth in this paper, three sample designs are briefly presented. The first consists of the design of a "wideband" differentiating filter whose magnitude characteristics are to be within 0.1 db out to 80% of the Nyquist limit. Because the magnitude characteristic is not piecewise constant, the bilinear z-form may not be used satisfactorily. Instead the standard z-transform is used. A guard filter is necessary to reduce folding error at all frequencies to within specifications. The guard used is a 0.1 db ripple Chebyshev low-pass filter of ninth order chosen to give 50 db loss at 120% of the Nyquist limit. The magnitude response of the resulting sampled filter is shown in Fig. 4.

---

[3] D. Slepian, H. O. Pollak, and H. J. Landau, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty," Parts I and II, Bell System Technical Journal, Vol. 40, No. 1, pp. 43-84, January 1961.

To design a sampled-data bandstop filter having equiripple passband magnitude characteristics and equiripple stopband loss the bilinear z-form is used with prewarping of the band edge frequencies being required. The result of the design is shown in Fig. 5. This result could not have been arrived at by the standard z-transform method.

To illustrate the nonrecursive filter design procedure, the "wideband" differentiating filter of the first example is to be obtained in nonrecursive form. To satisfy the 0.1 db ripple requirement $\omega_a \tau$ is set at 6.0 for the modified Bessel weighting function. The value of $\omega_a/2\pi$ is set at 1 kc to satisfy the bandwidth requirement. The resulting τ/T is then 9.55. Using these parameters the coefficients of the nonrecursive filter are then found. The frequency response of this filter is shown in Fig. 6. In comparing this design with that of Fig. 4 it is seen that both designs have approximately the same number of coefficients but that the nonrecursive filter has a constant phase slope equal to an integer number of samples and obtained at the price of the larger delay, 9T as compared to 3T for the recursive design.

Summary

In this paper the essential features of basically three different design methods for sampled data filters have been treated briefly. By discussion and through examples these methods have been shown to cover a wide variety of design problems. The choice of methods presented was not intended to be complete (as evidenced by the omission of least squares, orthogonal function, and other polynomial methods) but only representative of the range of procedures available.

Absent from the discussion has been mention of the details of the numerical analysis problems such as required coefficient accuracy, error control, and noise propagation problems which show themselves primarily in the higher order filter designs. Obtaining detailed insight into the computational aspects of these design methods is essential for their effective and intelligent use in realizing satisfactory simulations of continuous filter networks.

List of Figure Captions

Fig. 1	Nonlinear warping of the frequency scale in the bilinear z-transformation.
Fig. 2	Nonrecursive filter design procedure using weighting functions.
Fig. 3	A modified Bessel function of the first kind and zeroth order as a weighting function.
Fig. 4	The frequency response of a sampled data differentiating filter of the recursive type. The filter was obtained by the standard z-transform utilizing a cascade guard filter.
Fig. 5	Frequency response of a sampled data bandstop filter of the elliptic type obtained using the bilinear z-transformation with warping correction.
Fig. 6	Frequency response of a nonrecursive differentiating filter obtained using a modified Bessel function spectral window for truncation.

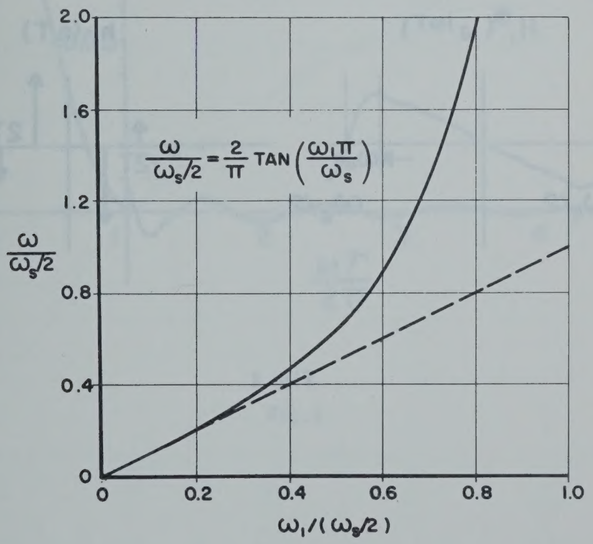


Fig.1

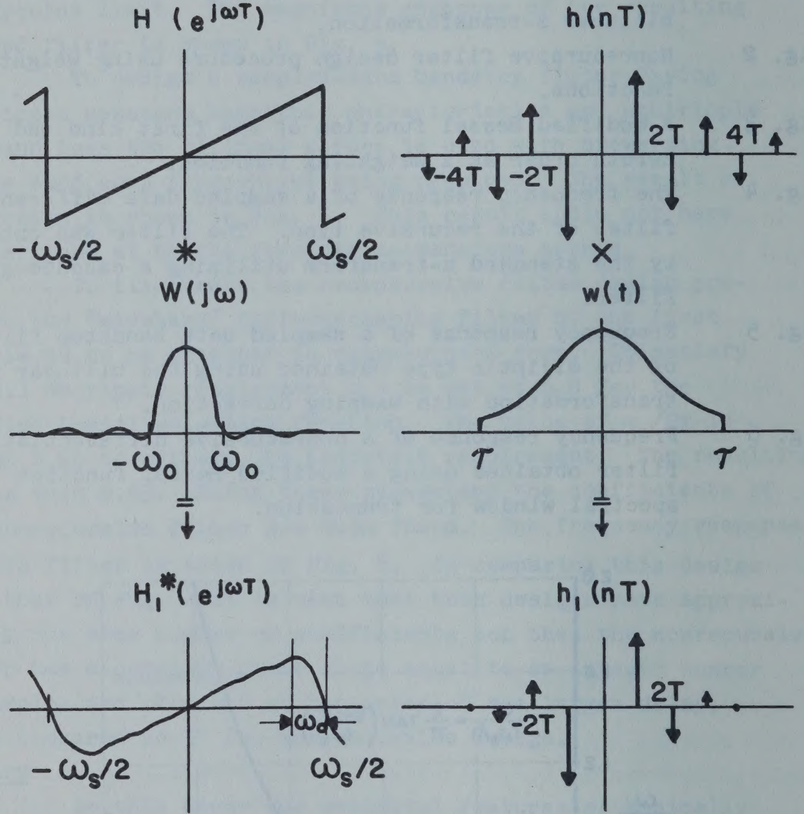

H (e^{jωT})
h(nT)
-ω_s/2
*
ω_s/2
-4T
-2T
2T
4T
×
W(jω)
w(t)
-ω_o
ω_o
τ
τ
=
=
H_1*(e^{jωT})
h_1(nT)
-ω_s/2
ω_d
ω_s/2
-2T
2T

Fig. 2

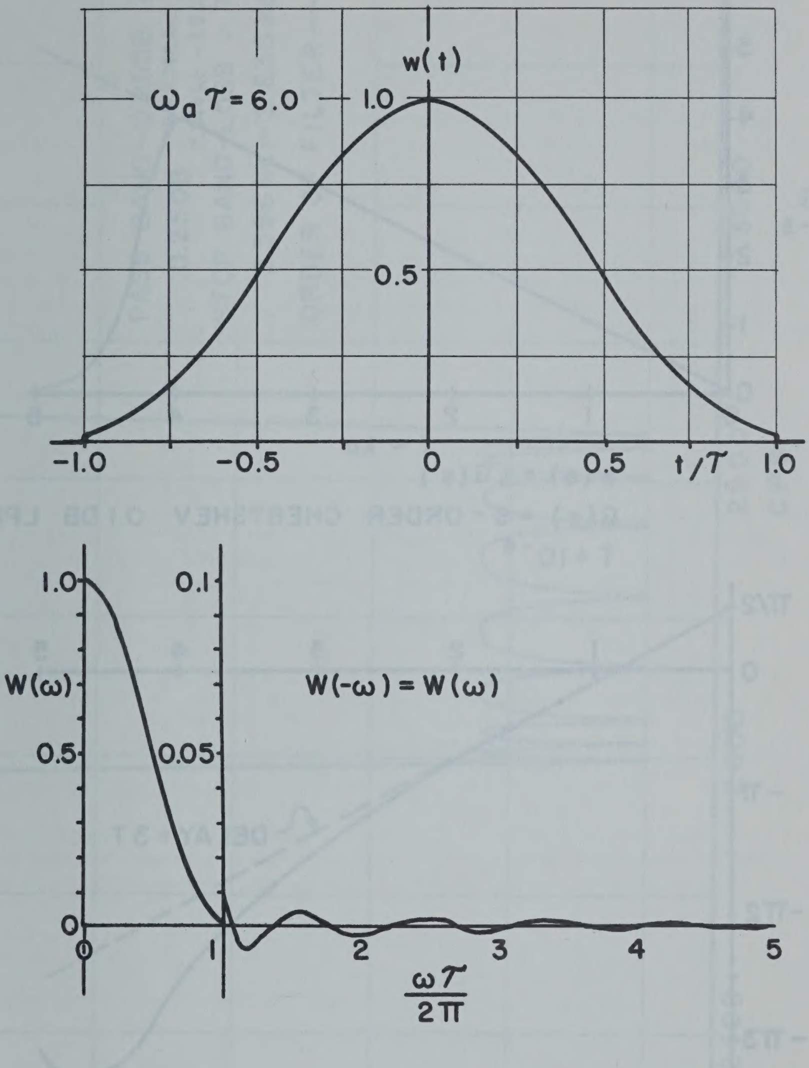


Fig. 3

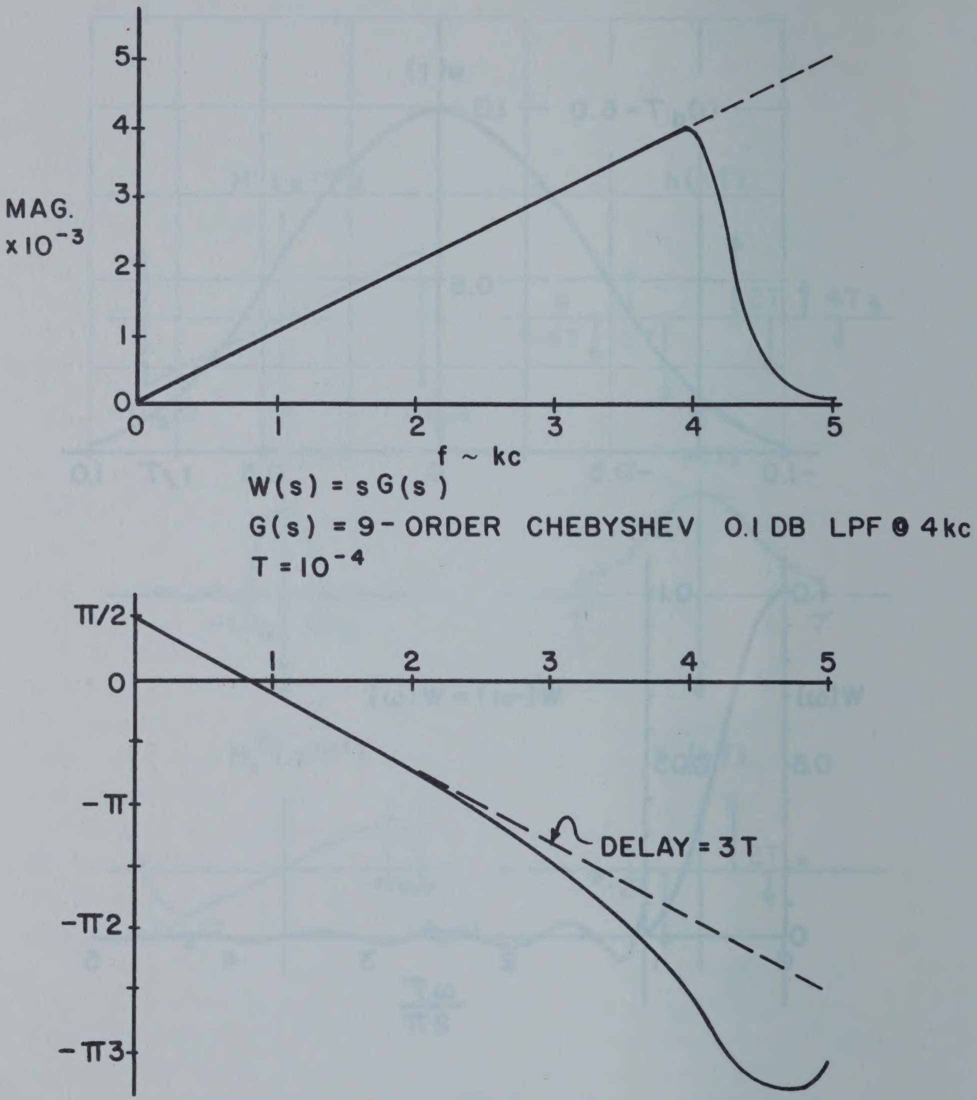

MAG.
x 10^{-3}
5
4
3
2
1
0
0
1
2
3
4
5
f ~ kc
W(s) = s G(s)
G(s) = 9 - ORDER CHEBYSHEV 0.1 DB LPF @ 4 kc
T = 10^{-4}
π/2
0
-π
-π2
-π3
1
2
3
4
5
DELAY = 3 T

Fig.4

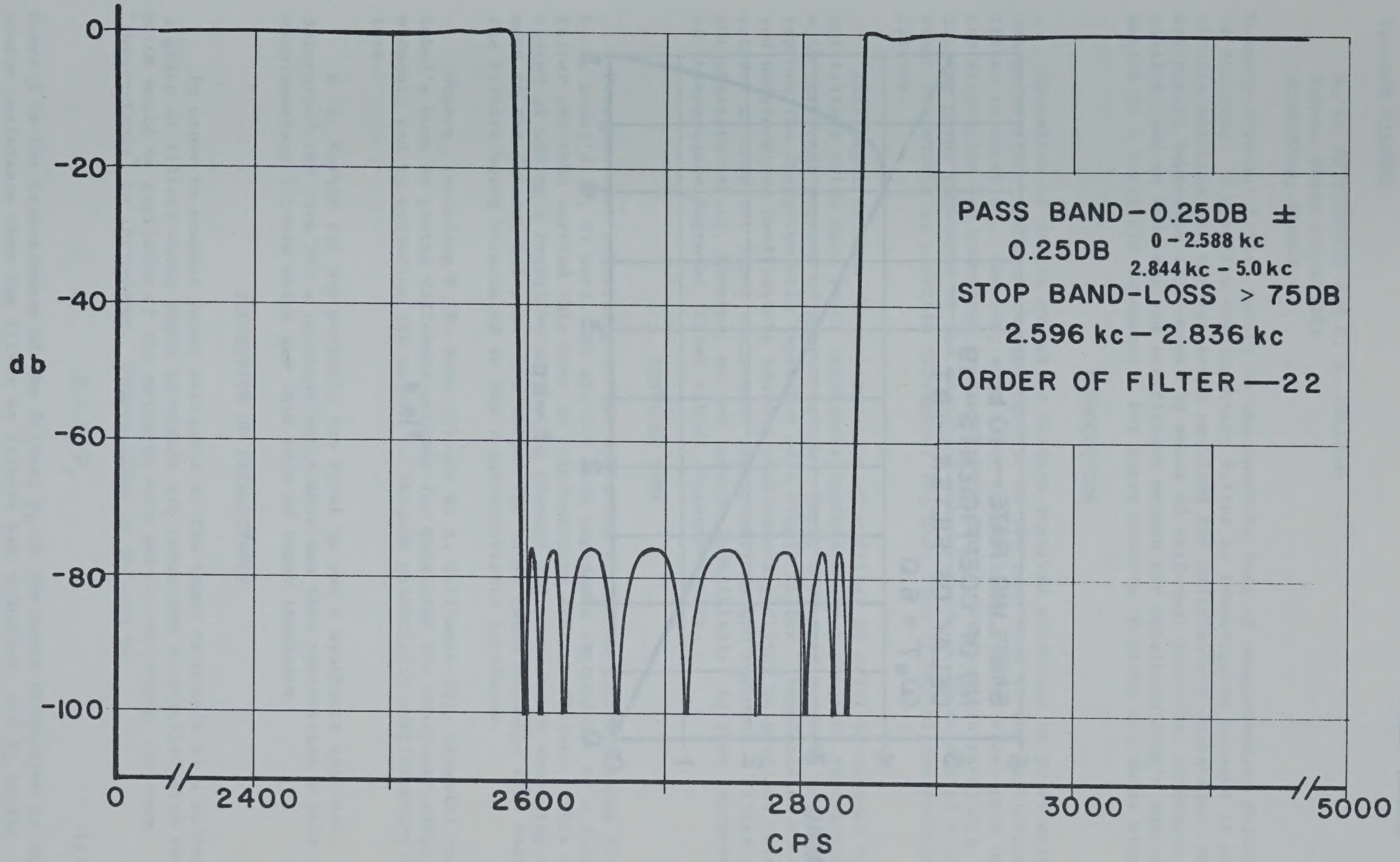

0
-20
-40
-60
-80
-100
db
PASS BAND-0.25DB ±
0.25DB
0 - 2.588 kc
2.844 kc - 5.0 kc
STOP BAND-LOSS > 75DB
2.596 kc — 2.836 kc
ORDER OF FILTER —22
0
2400
2600
2800
3000
5000
CPS

Fig.5

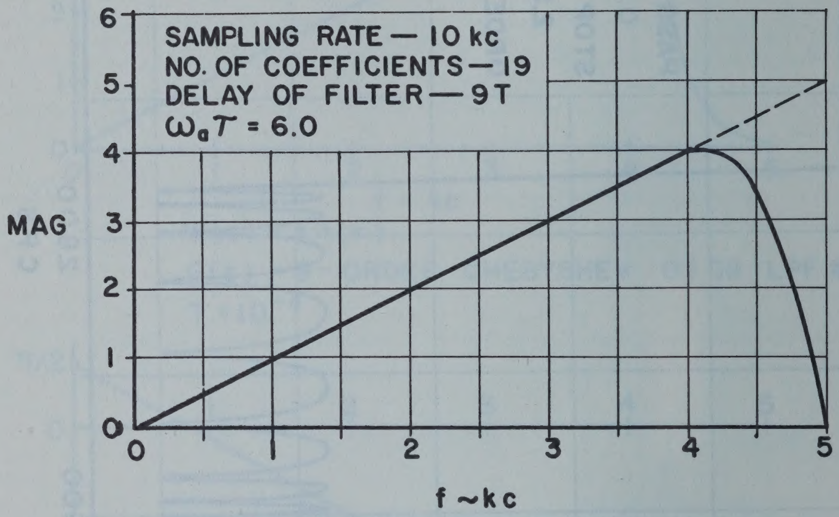

SAMPLING RATE — 10 kc
NO. OF COEFFICIENTS — 19
DELAY OF FILTER — 9T
$\omega_a T = 6.0$
MAG
0
1
2
3
4
5
6
0
1
2
3
4
5
f ~ kc

Fig. 6

FAN-OUT FILTERS

R. D. FRITZEMEYER and C. A. HALIJAK
Kansas State University
Manhattan, Kansas

Summary--Fan-out filter theory in the specific form of complementary filters is reviewed. A specific complementary filter is investigated because it succinctly exhibits the several steps required for satisfactory operation: characteristic impedance improvement by means of self-dual networks; impedance elision; and an undetermined coefficient method for obtaining King's approximation to a resistive impedance. Not every network requires all these steps.

INTRODUCTION

Operation of single symmetric filters requires attention be given only to characteristic impedance characteristics in the pass band of the particular filter in question. However, when input terminals of filters whose pass bands are dissimilar are connected in fan-fashion (input terminals in parallel), the driving point impedance of the resultant network needs to be as nearly a constant resistance as possible throughout the pass band of any of the individual filters.

This paper defines the interactance of a filter in order to compare fan-out filters which have input impedances approximating a constant resistance with different degrees of approximation. Next, self-dual networks are used to improve the characteristic impedance of a single filter. Impedance elision and undetermined coefficients improve complementary filter outputs and interactance as constant input resistance approximations are improved. Filter output characteristics, however, do not improve indefinitely; further improvement of interactance degrades filter output characteristics.

PREVIOUS WORK

Previous work pertaining to the operation of fan-out filters stems from O. J. Zobel's (1, 2) work. He started with impedance improvement of a single filter and then carried this over to complementary fan-out filters. This consisted of adding a resistive correcting element and a reactance annulling element to the fan-out filters. The values for these added elements were based on the filters being terminated in their characteristic impedance.

Others, including W. H. Bode (3) and E. A. Guillemin (4), expanded upon Zobel's work by giving different methods for obtaining the reactance annulling elements and by extending the method to include potentially complementary filters.

E. L. Norton (5) was probably the first to use a synthesis approach. He determined the form for a constant resistance and then synthesized a pair of complementary filters which had this form of input impedance.

DEFINITION OF INTERACTANCE

In order to compare power available at the input terminals of a filter or a group of filters whose input terminals are connected in parallel with that which would be available if the networks were purely resistive, the term "interactance" is introduced. Interactance is defined as

$$\Lambda = P_f/P_r \tag{1}$$

where Λ is the interactance of the filter, P_f is the power dissipated in the source resistance when the filter or filters are connected, and P_r is the power dissipated in the source resistance when a pure resistance replaces each filter in the circuit. These powers, P_f and P_r, are measured with an additional voltage put in each parallel filter path (see Fig. 1) such that when the filter is replaced by a pure resistance the voltage across the source resistance is equal to the original source voltage. Figure 2 shows the filters replaced by one-ohm (which is the pseudo-characteristic impedance of the filters) resistors. Using Millman's theorem, the voltage e_{or} in Fig. 2 is

$$e_{or} = \frac{1(e) + n(-e/n)}{1 + n} = 0 \qquad (2)$$

where n is the number of parallel paths. Therefore the voltage across the source resistance, e_r, is

$$e_r = e - e_{or} = e \qquad (3)$$

where e is the source voltage. The power, P_r, dissipated in the source resistance under these conditions is (assuming e is a peak value)

$$P_r = e^2/2R \qquad (4)$$

If R is one ohm, Eq. (4) reduces to

$$P_r = e^2/2 \qquad (5)$$

Calculation of P_f follows a similar procedure except the one-ohm load resistors are replaced by filter input admittances, Y_n, as in Fig. 1, and yields

$$e_{of} = \frac{Y_0(e) + Y_1 \frac{-e}{n} + Y_2 \frac{-e}{n} + \dots + Y_n \frac{-e}{n}}{Y_0 + Y_1 + Y_2 + \dots + Y_n} \qquad (6)$$

The voltage across the source resistance for this case is

$$e_r = e - e_{of} \qquad (7)$$

Substituting Eq. (6) into Eq. (7) yields

$$e_r = e \frac{\frac{n+1}{n}(Y_1 + Y_2 + \dots + Y_n)}{Y_0 + Y_1 + Y_2 + \dots + Y_n} \qquad (8)$$

The power, P_f, in this case is

$$P_f = \frac{1}{2} e^2 Y_0 {\frac{\frac{n+1}{n}(Y_1 + Y_2 + \dots + Y_n)}{Y_0 + Y_1 + Y_2 + \dots + Y_n}}^{2} \qquad (9)$$

Substituting Eqs. (5) and (9) into Eq. (1), the interactance is found to be

$$\Lambda = Y_0 \frac{\frac{n+1}{n}(Y_1 + Y_2 + . . . + Y_n{}^{2}}{Y_0 + Y_1 + Y_2 + . . . + Y_n} \tag{10}$$

in terms of the input admittances of the individual filters in the circuit.

After calculating the interactance, it is noted that the plot of ideal interactance versus frequency is a straight horizontal line. The value for the ideal interactance is one. However, as in the case of complementary filters, the admittance of one filter may be zero so that the added voltage in that branch adds no power to the circuit. Therefore the best interactance plot that may be obtained for this case is a straight line of some value less than one. Thus it is necessary to consider each case separately in order to determine the best value of interactance obtainable, although the ideal straight line applies to all cases.

INTERACTANCE OF INDIVIDUAL FILTERS

Constant-K "T" Filter

Considered individually, the symmetrical "T" or "π" constant-K filter has an input impedance which approximates a pure resistance of K ohms. This is seen by examining the input impedance of a low-pass "T" constant-K filter. The expression for the input impedance (K normalized to one ohm) is

$$Z_{in} = \frac{1 + 2s + 2\lambda s^2 + 2(2 - \lambda)s^3}{1 + 2s + 2(2 - \lambda)s^2} \tag{11}$$

If λ is chosen such that the maximum number of coefficients of corresponding powers of s in the numerator and denominator are equal, the filter is symmetrical and the input impedance is

$$Z_{in} = \frac{1 + 2s + 2s^2 + 2s^3}{1 + 2s + 2s^2} \tag{12}$$

This is King's (6) approximation to a constant for small values of s--as many as possible of the corresponding coefficients of similar powers of s in the numerator and denominator polynomials are pairwise equal.

The interactance of this low-pass filter is calculated by use of Eq. (10) and is

$$\Lambda = \frac{1 + 2s + 2s^2}{1 + 2s + 2s^2 + s^3}{}^{2}_{s=j\omega} \tag{13}$$

Evaluating Λ for $s = j\omega$, the following result is obtained.

$$\Lambda = \frac{1 + 4\omega^4}{1 + \omega^6} \tag{14}$$

This shows that the interactance for this filter would be a constant only for small values of ω.

M-Derived Filter

The m-derivation of a constant-K filter, which is included in most network textbooks (7,8), is one means of improving the impedance characteristics throughout the pass band of the filter. It also makes it possible to choose one frequency of infinite attenuation which depends upon the value of m that is chosen. This means that it not only has a flatter impedance characteristic in the pass band but also has a sharper cut-off at the start of the attenuation band. Therefore the interactance is flat throughout more of the pass band than was the case with the constant-K filters. However, the infinite attenuation outside the pass band causes trouble when a high- and a low-pass m-derived filter are connected in fan fashion. Thus methods other than characteristic impedance improvement are required for ideal operation of fan-out filters.

INTERACTANCE OF COMPLEMENTARY FAN-OUT FILTERS

Fan-out Constant-K Filters

If two complementary constant-K filters are connected in fan fashion, the interactance of the combined network is calculated to be

$$\Lambda = \left. \frac{3/2(1 + 4s + 8s^2 + 9s^3 + 8s^4 + 4s^5 + s^6)^2}{2(1 + 3.5s + 6.5s^2 + 7.75s^3 + 6.5s^4 + 3.5s^5 + s^6} \right|_{s=j\omega} \tag{15}$$

That the equivalent input impedance for the fan-out constant-K filters is a ratio of polynomials of higher degree of s than for each filter by itself suggests that more of the coefficients of corresponding powers of s in the numerator and denominator polynomials could be made equal and the plot of interactance could be a much flatter curve. One method of accomplishing this is to write the expression for the equivalent input impedance with a variable coefficient for the reactance nearest the input terminals in each filter. This gives an equivalent input impedance of

$$Z_{in} = \frac{A_0 + A_1 s + A_2 s^2 + A_3 s^3 + A_4 s^4 + A_5 s^5 + A_6 s^6}{B_0 + B_1 s + B_2 s^2 + B_3 s^3 + B_4 s^4 + B_5 s^5 + B_6 s^6} \tag{16}$$

where

$$\begin{aligned} A_0 &= 2k_1 & B_0 &= 2k_1 \\ A_1 &= 2k_1(k + 2) & B_1 &= 6k_1 + 2 \\ A_2 &= 6kk_1 + 3k + 1 & B_2 &= 9k_1 + 2k + 5 \\ A_3 &= 10kk_1 + 2k + 1 & B_3 &= 6k_1 + 6k + 6 \\ A_4 &= 6kk_1 + 3k + 1 & B_4 &= 2k_1 + 9k + 5 \\ A_5 &= 2k(k_1 + 2) & B_5 &= 6k + 2 \\ A_6 &= 2k & B_6 &= 2k \end{aligned} \tag{17}$$

Choose k and k_1 so that the input impedance is as close an approximation to one, in both pass bands, as can be obtained. Since the coefficient pairs of the zero and sixth powers of s will be equal for any value of k and k_1, the coefficients of the first power of s and the fifth power of s are examined. The coefficient of s in the denominator is set equal to the numerator s coefficient. The same procedure is followed for the coefficients of the fifth power of s. This yields

$$k = k_1 = 1.618 = \frac{1 + 5}{2} \tag{18}$$

If the magnitude of the voltage transfer function of each of the filters is plotted versus frequency for the filters by themselves and in parallel, it is seen that there is a big improvement when their inputs are paralleled and the first series impedance changed to 1.618 times its original value. In fact, it approaches the flatness and sharpness of the cut-off of the transfer characteristic of a single m-derived filter.

The procedure for improving the input impedance characteristics can be carried further to obtain fan-out filters with a constant input impedance. However, the voltage transfer function is not as flat for this case. Therefore some compromise must be made in a practical situation between the input impedance and the filter characteristics required. For lengthy improvement details see Fritzemeyer (9). This constitutes empirical detection of a constraint--that filter improvement does not necessarily follow interactance improvement.

Fan-out M-derived Filters

The same approach can be used to improve the interactance for fan-out m-derived filters as was used with the constant-K filters. However, if a shunt branch is nearest the input of the m-derived filter, it must be omitted from the circuit when used in fan fashion. This will be called the degenerate m-derived case. The interactance of these filters is somewhat better than the constant-K filters and also has a better voltage transfer characteristic.

Self-dual Filter for Input Impedance Improvement

The input impedance of a low-pass symmetrical constant-K "T" filter is

$$Z = \frac{1 + 2s + 2s^2 + 2s^3}{1 + 2s + 2s^2} \tag{19}$$

If the impedance level is doubled, the input impedance is

$$Z_{in} = \frac{2 + 4s + 4s^2 + 4s^3}{1 + 2s + 2s^2} \tag{20}$$

which may also be written as

$$Z_{in} = 2 + \frac{4s^3}{1 + 2s + 2s^2} = 2 + \epsilon \tag{21}$$

The input impedance of a dual filter with the impedance level doubled is

$$Z_{in} = \frac{2 + 4s + 4s^2}{1 + 2s + 2s^2 + 2s^3}$$

$$= 2 - \frac{2s^3}{1 + 2s + 2s^2 + 2s^3} \simeq 2 - \epsilon \tag{22}$$

If these two filters are connected so that their inputs are in parallel, the input impedance is a closer approximation to a resistance for small values of ϵ.

$$\frac{(2 + \epsilon)(2 - \epsilon)}{2 + \epsilon + 2 - \epsilon} = 1 - \frac{\epsilon^2}{4} \cong 1 \qquad (23)$$

Since the output voltage of each filter is the same, the output terminals may also be connected in parallel. The resulting network is the self-dual network shown in Fig. 3. Reduction of this network to an equivalent ladder structure yields the low-pass filter of Fig. 4.

Fan-out Operation of Self-dual Filters

A high-pass filter with an improved impedance characteristic may be obtained in a similar manner. Fan-out operation of these two filters is not very satisfactory since the low-pass filter has an impedance zero at $\omega = \infty$ and the high-pass filter has an impedance zero at $\omega = 0$. However, if the parallel branch nearest the input terminals of each of the filters is removed, as in the fan-out operation of m-derived filters, the input impedance resistive characteristic greatly improved. Figure 5 shows the interactance of these networks which is an improvement over the constant-K filter interactance. Also the voltage transfer functions (see Fig. 6) of these filters connected in fan fashion are an improvement over the transfer functions of the constant-K filters. In fact, they have similar cut-off characteristics to the transfer functions of the degenerate m-derived filters.

The interactance of complementary fan-out filters derived from the parallel connection of dual filters may be improved even more by the method used on the constant-K and the m-derived filters. This involves changing of the first series element in each filter to obtain the coefficient of s^1 in the denominator of the input impedance function. Carrying out these calculations gives a value of 1.167s instead of s as shown in Fig. 4. Figure 7 shows the improved interactance curve. As was the case with the constant-K and m-derived filters, this process of changing element values could be carried to such an extent as to yield complementary filters with a constant resistance input impedance. The same constraint was present also.

CONCLUSIONS

A constant-K complementary filter theory has been exhibited. It deals with King's approximation to a resistive driving point impedance as contrasted to Zobel's and Norton's use of characteristic impedances. This is an indirect filter theory because the outputs are not pre-specified in detail. This procedure can be carried out only a finite number of times--a fact which has been empirically detected in the course of this investigation.

Future investigations are suggested by these efforts. The 3-db point is not preserved and an investigation of the Guillemin (4) problem is in order. Spectrum analyzers require large numbers of filter outputs and it is apparent that a constant-K spectrum analyzer will have 2^n outputs (where n is the number of stages in the tree).

REFERENCES

1. Zobel, O. J.
   U.S. Patent 1,557,230.
2. Zobel, O. J.
   Distortion correction in electrical circuits with constant resistance recurrent networks. Bell System Technical Journal, July, 1928, pp. 438-534.
3. Bode, H. W.
   A method of impedance correction. Bell System Technical Journal, October, 1930, pp. 794-835.

4. Guillemin, E. A.
Communication networks. New York: John Wiley and Sons, Vol. 2, 1935.

5. Norton, E. L.
Constant resistance networks and application to filter systems. Bell System Technical Journal, Vol. 16, 1937, pp. 178-193.

6. King, L. H.
Reduction of forced error in closed-loop systems. Proc. I.R.E., August, 1953, pp. 1037-1042.

7. Karakash, J. J.
Transmission lines and filter networks. New York: The Macmillan Company, 1950.

8. Starr, A. T.
Electric circuits and wave filters. New York: Pitman Publishing Corporation, 1938.

9. Fritzemeyer, R. D.
Fan-out filters. M.S. Thesis, Dept. of Electrical Engineering, Kansas State University; Kansas State Engineering Experiment Station Special Report No. 36, June, 1963.

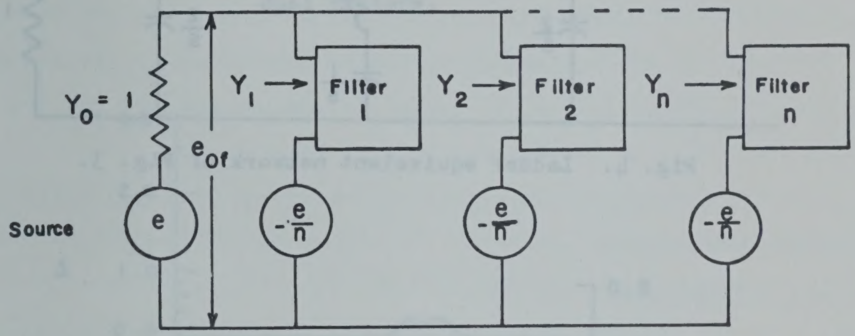

$Y_0 = 1$
Y_1
Filter 1
Y_2
Filter 2
Y_n
Filter n
e_{of}
Source
e
$-\frac{e}{n}$
$-\frac{e}{n}$
$-\frac{e}{n}$

Fig. 1. The circuit arrangement for measurement of the interactance of a group of filters whose inputs are in parallel.

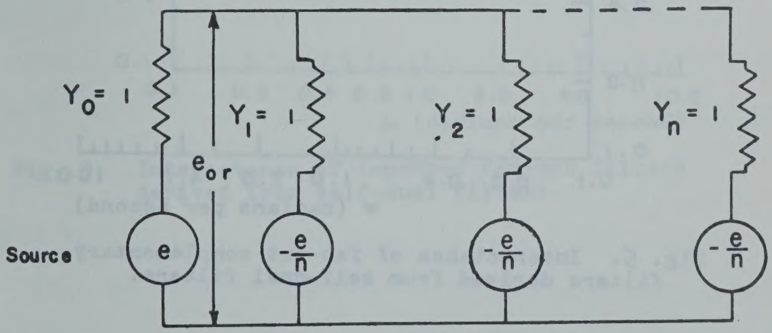

$Y_0 = 1$
$Y_1 = 1$
$Y_2 = 1$
$Y_n = 1$
e_{or}
Source
e
$-\frac{e}{n}$
$-\frac{e}{n}$
$-\frac{e}{n}$

Fig. 2. The circuit arrangement for measurement of the interactance with the filters replaced by one-ohm resistors.

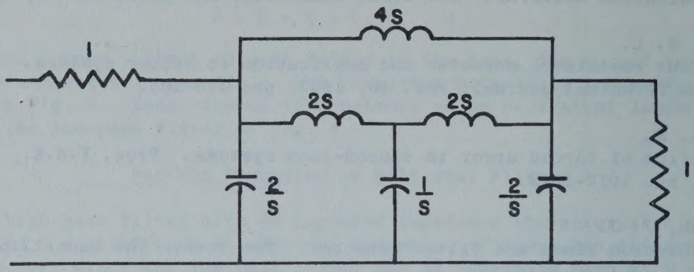

4S
1
2S
2S
2/S
1/S
2/S
1

Fig. 3. The self-dual network resulting from the paralleling of dual low-pass constant-K filters.

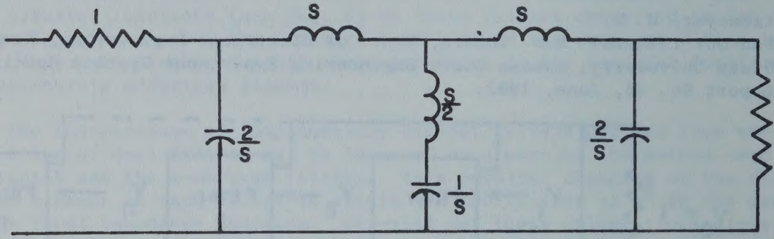

1
S
S
S/2
2/S
1/S
2/S
1

Fig. 4. Ladder equivalent network of Fig. 3.

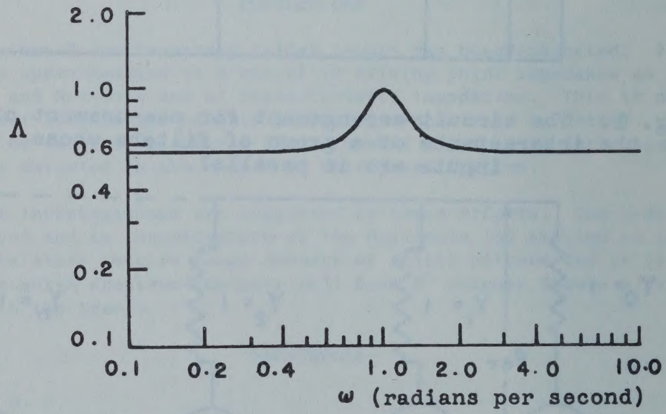


Fig. 5. Interactance of fan-out complementary filters derived from self-dual filters.

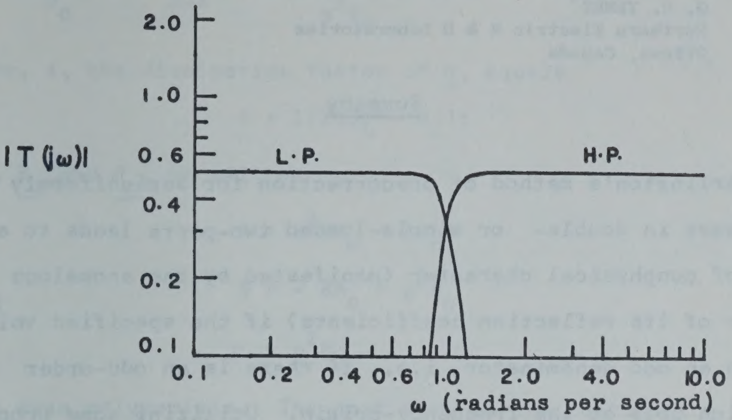


Fig. 6. Voltage transfer functions of fan-out complementary filters derived from self-dual filters.

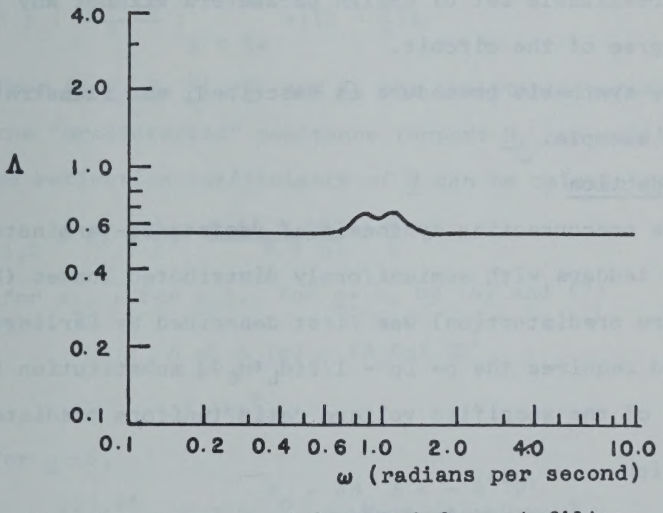


Fig. 7. Interactance of improved fan-out filters derived from self-dual filters.

AN EXTENSION OF DARLINGTON'S SEMIUNIFORM PREDISTORTION PROCEDURE

G. C. TEMES
Northern Electric R & D Laboratories
Ottawa, Canada

Summary

Darlington's method of precorrection for semiuniformly distributed losses in double- or single-loaded two-ports leads to a network of nonphysical character (manifested by the anomalous behaviour of its reflection coefficients) if the specified voltage ratio has an odd denominator, i.e., if there is an odd-order attenuation pole at the frequency-origin. Utilizing some properties of networks containing oppositely dissipative reactances, a predistortion procedure, analogous to Darlington's, was developed for networks synthesized from such voltage ratios. The process yields a realizable set of design parameters without any increase in the degree of the circuit.

The synthesis procedure is described, and illustrated by a numerical example.

I. Introduction

The precorrecting synthesis of resistance-terminated reactance ladders with semiuniformly distributed losses (briefly: semiuniform predistortion) was first described by Darlington [1]. His method requires the $p \rightarrow [p - 1/2(d_L + d_C)]$ substitution in the numerator of the specified voltage ratio (uniform predistortion). The function

$$\Lambda = \frac{A_o + pB_o}{P_o} \qquad (1)$$

thus obtained must next be realized as the voltage ratio of a network N containing oppositely dissipative coils and capacitors, with dissipation constants δ and $-\delta$, respectively. To this end, Λ is brought (by including surplus factors) into a special form:

$$\Lambda = \frac{A_o + pB_o}{P_o} = \frac{p}{p-\delta} \; \frac{p(p-\delta)(A_o + pB_o)}{p^2 P_o} = \frac{p}{p-\delta} \; \frac{A + pB}{P} \qquad (2)$$

Here, δ, the dissipation factor of N, equals

$$\delta = 1/2(d_L - d_C); \qquad (3)$$

A_o, B_o and P_o, as well as

$$A = p^2(A_o - \delta B_o) \qquad (4)$$

and

$$B = -\delta A_o + p^2 B_o \qquad (5)$$

$$P = p^2 P_o \qquad (6)$$

are even polynomials. The next step is the calculation of two additional even polynomials, A' and B', from

$$A'^2 - p^2 B'^2 = A^2 - p^2 B - K^2 P^2 \qquad (7)$$

where K is a constant, subject to the condition

$$K \leq \left| \frac{A + pB}{P} \right|_{p = j\omega} = \left| (1 - \frac{\delta}{p}) \Lambda \right|_{p = j\omega} \qquad (8)$$

From A, B, P, A', B' and K, the immittance parameters of N and of the "predistorted" reactance twoport N_R are easily obtained. Also, the reflection coefficients of N can be calculated from

$$\rho_{1,2} = (1 - \frac{\delta}{p}) \; \frac{\pm A' + pB'}{A + pB} - \frac{\delta}{p} \qquad (9)$$

(+ sign for ρ_1, - for ρ_2). For $p \to o$, by (4) and (7)

$$A \to [\, A_o(o) - \delta B_o(o) \,] p^2 \qquad (10)$$

and

$$A' \to kp^2 . \qquad (11)$$

Hence, for $p \to 0$,

$$\rho_{1,2} \to \left[\frac{1}{p} \left(\frac{B'}{A_o} - \delta \right) + \frac{A_o - \delta B_o \pm k - \delta^{-1} B'}{A_o} \right]_{p \to o} \qquad (12)$$

Now unless

$$\frac{B'(o)}{A_o(o)} = \delta, \qquad (13)$$

both $\rho_{1,2} \to \infty$, i.e., the driving-point impedances Z_1 and Z_2 of N become equal to $-R_1$ and $-R_2$, respectively, *regardless of the value of δ.*

This is a physically inconceivable condition, and it leads to unrealizable network functions. Fortunately, (13) can always be satisfied, since, by (7)

$$B'(o) = B(o) = \mp \delta A_o. \tag{14}$$

Hence, physical solution may be obtained by selecting the sign of B' "properly", i.e., according to (13).

Suppose now the P_o is odd. Then

$$\frac{A_o + pB_o}{P_o} = \frac{p}{p-\delta} \; \frac{(p-\delta)\,(A_o + pB_o)}{pP_o} = \frac{p}{p-\delta} \; \frac{A + pB}{P} \; . \tag{15}$$

Here,

$$A = -\delta A_o + p^2 B_o \tag{16}$$

$$B = A_o - \delta B_o \tag{17}$$

$$P = pP_o \tag{18}$$

are again all even. Eqs. (7) - (9) will still be valid and hence, for $p \to o$, from (7), (9) and (16) - (17),

$$\lim_{p\to o} \rho_{1,2} = \left[\frac{1}{p} \left(\frac{\pm A'}{A_o} - \delta \right) + \frac{B + B' \mp A'/\delta}{A_o} \right]_{p\to o} . \tag{19}$$

Hence, for any sign of A', one of the reflection coefficients becomes infinite[1]. (Similar difficulties occur for single-loaded networks). Therefore, Darlington's procedure is not directly applicable to voltage ratios with odd denominators. In the following a method, analogous to Darlington's, will be given for handling such voltage ratios.

II. The Voltage Ratio of Oppositely Dissipative Twoports

Consider the reactance twoport N_R obtained by removing the losses from the elements of N. The immittance-parameters of this twoport have either the form $\frac{\bar{f}_{ij}}{p}$ or $p\bar{f}_{ij}$, where $\bar{f}_{ij}$ is an even rational function of p. Straightforward circuit analysis, based on the expression of the immittance-parameters with the circuit-determinant and its minors, shows that by adding the ±δ losses these parameters are changed according to Table I.

Parameter	Reactive Network (N_R)	Oppositely Lossy Network (N)
Z_{ij}	$\frac{\bar{f}_{ij}}{p}$	$\frac{f_{ij}}{p-\delta}$
	$p\bar{f}_{ij}$	$(p+\delta)\, f_{ij}$
Y_{ij}	$\frac{\bar{g}_{ij}}{p}$	$\frac{g_{ij}}{p+\delta}$
	$p\bar{g}_{ij}$	$(p-\delta)\, g_{ij}$

TABLE I

The convention adopted in Table I (and in the rest of this paper) is that even functions with and without bar are related by

$$\begin{aligned} f(p^2) &= \bar{f}(p^2 - \delta^2) \\ \bar{f}(p^2) &= f(p^2 + \delta^2) \end{aligned} \tag{20}$$

A simple analysis based on the reactive character of N_R and on the results of Table I shows that while the voltage ratio of N, Λ, may have an *even*-order pole at $p = o$ (if the corresponding N_R has an attenuation pole at $p = \pm j\delta$), it cannot have an *odd*-order pole there. If N_R has a $(2n+1)$th -order attenuation pole at the frequency-origin, Λ will have a nth-order pole at $p = \pm\delta$ and a single pole at either $p = \delta$ or $p = -\delta$, depending upon the forms of the immittance-parameters (Table I). It is also evident that $p = \delta$ or $-\delta$ are the *only* possible locations for an odd-order pole of Λ, if it is to lead to a realizable N_R.[3] Hence, the proper form of Λ is

$$\Lambda = \frac{A + pB}{(p \pm \delta)\, P'} \tag{21}$$

where, A, B and P are again even polynomials[2].

The main difference between (21) and Darlington's expression (Eq. (2) in this paper) is that the Λ function of Eq. (21) does not contain surplus factors and hence $p = \delta$ or $-\delta$ is in fact an attenuation pole of N. The proper sign of δ can readily be found from the desired configuration. In the asymptotic circuit diagram o N_R, drawn for $p \to o$,[3] the number of shunt coils will be (n+1) and the number of series capacitors n, or vice versa. In the former case, the pole will be located at $p = -\delta$; in the latter, at $p = +\delta$, as a consideration of the dependence of Λ upon the branch immittances immediately shows.

III. The Synthesis of Double-Loaded Twoports

The insertion voltage ratio and its dual parameter, the insertion current ratio, are equal (apart from perhaps a difference in sign, occuring if the positive directions chosen for V_2 and I_2 conflict). Also, the dual network of an oppositely dissipative circuit is again an oppositely lossy network, in which, however, the dissipation factor is δ for the capacitors and $-\delta$ for the coils. From these statements it can be concluded that if an oppositely lossy network is synthesized from a voltage ratio $\Lambda(p,\delta)$, the network realizing $\pm\Lambda(p,-\delta)$ is immediately available. For this reason it is sufficient to treat the synthesis from voltage ratios of the form

$$\Lambda = \frac{A + pB}{(p - \delta)\, P} \tag{22}$$

This expression will result from impedance-parameters in the form

$$\frac{Z_{ij}}{\sqrt{R_i R_j}} = \frac{f_{ij}}{p-\delta} \tag{23}$$

or, equivalently, from admittance-parameters in the form

$$\frac{Y_{ij}}{\sqrt{G_i G_j}} = (p-\delta)\, g_{ij} \tag{24}$$

Expressing Λ with the Z_{ij} and using (22) - (23),

$$\frac{A + pB}{P} = \frac{K}{2}\, \frac{(p-\delta)(f_{11} + f_{22}) + |f| + (p-\delta)^2}{f_{12}} \tag{25}$$

results, where $K = 2\dfrac{\sqrt{R_1 R_2}}{R_1 + R_2}$ and

$$|f| = f_{11} f_{22} - f_{12}^2 . \tag{26}$$

Equating even and odd parts,

$$\frac{p^2 + (f_{11} - \delta)(f_{22} - \delta) - f_{12}^2}{f_{12}} = \frac{2}{K}\frac{A}{P} \tag{27}$$

and

$$\frac{(f_{11} - \delta) + (f_{22} - \delta)}{f_{12}} = \frac{2}{K}\frac{B}{P} \tag{28}$$

It is fairly evident that the introduction of the even polynomials A' and B', defined by

$$\frac{p^2 - (f_{11} - \delta)(f_{22} - \delta) + f_{12}^2}{f_{12}} = \frac{2}{K}\frac{A'}{P} \tag{29}$$

and

$$\frac{(f_{22} - \delta) - (f_{11} - \delta)}{f_{12}} = \frac{2}{K}\frac{B'}{P}, \tag{30}$$

respectively, is expedient. From (27) - (30), using (23)

$$Z_{11} = \frac{R_1}{p-\delta} \quad \frac{B - B' + \delta D}{D} \tag{31}$$

$$Z_{12} = \frac{\sqrt{R_1 R_2}}{p-\delta} \quad \frac{KP}{D} \tag{32}$$

$$Z_{22} = \frac{R_2}{p-\delta} \quad \frac{B + B' + \delta D}{D} \tag{33}$$

where

$$D = \frac{A + A'}{p^2}. \tag{34}$$

From (26) - (30)

$$A'^2 - p^2 B'^2 = A^2 - p^2 B^2 + K^2 p^2 P^2 \tag{35}$$

(cf. Eq. (7)). Since by their definition A' and B' are real polynomials,

$$\left| \frac{A' + pB'}{pP} \right|^2_{p=j\omega} \geqslant 0, \tag{36}$$

and from (35) - (36), using also (22)

$$K \leqslant \left| (1 - \frac{\delta}{p}) \Lambda \right|_{p=j\omega} \tag{37}$$

as in (8). (For minimum flat insertion loss, K should be chosen as large as (37) permits.)

From (31) - (34), the reflection coefficients of N turn out to be

$$\rho_{1,2} = \left(1 - \frac{\delta}{p}\right) \frac{A' \pm pB'}{A + pB} - \frac{\delta}{p} \; . \tag{38}$$

For $p \to o$,

$$\rho_{1,2} \to \left[\frac{A' - \delta(B \pm B')}{A} - \frac{\delta}{p} \; \frac{A + A'}{A}\right]_{p \to o} . \tag{39}$$

Hence, unless

$$A'(o) = - A(o), \tag{40}$$

both reflection coefficients become infinite. Condition (40) can be satisfied since it does not contradict (35). Because of (40), D, defined in (34), is a polynomial.

From (35), taking into account (37) and (40), A' and B' can be calculated. Then (31) - (34) give the Z_{ij}.

Using Table I, the impedance-parameters, Z_{ijR} of N_R can easily be obtained from those of N. A consideration of the degrees of polynomials B, B', P and D shows then that the degrees of the impedances and hence the order of the network are the same as in the case of a lossless twoport synthesized from the same specified voltage ratio. (This was not considered possible by Ming [2] for a network with an odd-order attenuation pole at zero frequency).

From the immittance-parameters of N_R, its voltage-ratio and reflection coefficients can also be obtained:

$$\Lambda_R = \frac{\bar{A} + (p + \delta)\,\bar{B} + p\delta\bar{D}}{p\bar{P}} \tag{41}$$

$$\rho_{1,2_R} = \frac{-\bar{A} - \delta\bar{B} \pm p\bar{B}' + p^2\bar{D}}{\bar{A} + (p + \delta)\bar{B} + p\delta\bar{D}} \; . \tag{42}$$

As mentioned above, the network realizing Λ with a pole at $p = - \delta$ can be obtained from the dual equations, with the sign of δ changed. Hence, e.g., (31) will be replaced by

$$Y_{11} = \frac{G_1}{p+\delta} \frac{B - B' - \delta D}{D} \qquad (43)$$

etc. Table I can again be used to obtain the admittance-parameters of the predistorted reactance network N_R.

For bandpass filters, the termination ratio can be found by developing the network N_R from both sides and taking the ratio of corresponding elements [3]. For highpass filters it is also possible to calculate R_1/R_2 from

$$\lim_{p\to\infty} \frac{Z_{11_R}}{Z_{22_R}} = 1 \qquad (44)$$

Eq. (44) yields

$$\frac{R_1}{R_2} = \lim_{p\to\infty} \frac{\bar{B} + \bar{B}'}{\bar{B} - \bar{B}'} \qquad (45)$$

for networks with a pole at $p = \delta$, and the reciprocal value for networks with a pole at $p = -\delta$.

IV. The Synthesis of Single-Loaded Networks

The calculations are considerably simplified for twoports which are resistively terminated at one port only. The procedure is now illustrated for a network with an odd-order pole at $p = -\delta$. The assumed terminations are shown in Fig. 1. The duality principle, the reciprocity theorem as well as Thevenin and Norton equivalences can be used to design networks with a pole at $p = +\delta$ and/or different terminations.

From Table I and from the expression of Λ in terms of the Z_{ij}, it is obvious that the form

$$\frac{Z_{ij}}{R_1} = (p + \delta)\, f_{ij} \qquad (46)$$

is now useful. Then

$$\Lambda = \frac{A + pB}{(p + \delta)P} = \frac{(p + \delta) f_{11} + R_1}{(p + \delta)\, f_{12}} \quad . \qquad (47)$$

From (46) - (47) it follows directly that

$$Z_{11} = R_1 \frac{(p + \delta)B}{A - \delta B} \tag{48}$$

$$Z_{12} = R_1 \frac{(p + \delta)P}{A - \delta B} \quad . \tag{49}$$

The Z_{ij_R} can be found using Table I. From Z_{ij_R}, the other design parameters of N_R turn out to be

$$\Lambda_R = \frac{\bar{A} + (p - \delta)\bar{B}}{p\bar{P}} \tag{50}$$

$$\rho_{1_R} = \frac{\bar{A} - (p + \delta)\,\bar{B}}{\bar{A} - (p - \delta)\,\bar{B}} \quad . \tag{51}$$

In case of short-circuited output terminals the dual equations should be used, with $\delta \to -\delta$. Thus, e.g., instead of (48), one has

$$Y_{11} = G_1 \frac{(p - \delta)\,B}{A + \delta B} \quad . \tag{52}$$

V. Numerical Example

The synthesis procedure is illustrated by the design of predistorted Chebyshev passband, general stopband bandpass filter. The specifications are the following:

Passband: $1 < f < 10$
Passband ripple: $A_p \leq 0{\cdot}3$ db
Terminations: $R_1 = 1,\ R_2 \geq 1$
Quality factors: $Q_C >> Q_L = 100$
Attenuations poles: $f_o = 0$ (single pole)
$f_1 = 11$
$f_2 = 12$
$f_\infty \to \infty$ (single pole) .

The characteristic function satisfying these specifications on a lossless basis is

$$\Phi = \frac{p^6 + 16.544886452463157615476233\ p^4 + 68.912491941396763433780470\ p^2 + 24.867958324338982982146063}{p(1.2500750253939265892068645 p^4 + 33.126988172939054613981909\ p^2 + 217.81307242463776890340407)} \tag{53}$$

The corresponding lossless voltage ratio can be calculated using (7). Uniform predistortion of the lossless voltage ratio yields

$$\begin{aligned} A + pB = {} & p^6 + 3.9073944755376488140428601p^5 + 23.417020434015101574243825p^4 \\ & + 62.322166866457432299313803p^3 + 134.18490715443503291067556p^2 \\ & + 223.89908124324976551956591p + 23.745100490525150967550928 \end{aligned} \qquad (54)$$

From (37), with some safety margin,

$$K^2 = 0.880846724635 \qquad (55)$$

was chosen. Then, from (35) and (40),

$$\begin{aligned} A' + pB' = - (p^6 & + 1.8874424221966168678370131p^5 \\ & + 18.252618202278532732602455p^4 \\ & + 26.30179165\ 4681814486026648p^3 \\ & + 84.387059382554315120616237p^2 \\ & + 77.306452789942645444101550p \\ & + 23.745100490525150967550921) \end{aligned} \qquad (56)$$

and from (34),

$$D = 5.1644022317365688416413 7p^2 + 49.797847771880718790059 32 \qquad (57)$$

The corresponding functions $\bar{A}$, $\bar{B}$, $\bar{B}'$ and $\bar{D}$ were obtained using (20). Anticipating the desired circuit configuration (Fig. 2), the admittance (rather than the impedance) parameters of $N_{\underline{R}}$ were calculated. Using (31) - (34),

$$R_i Y_{ii_R} = p \frac{\bar{B} \pm \bar{B}' + \delta \bar{D}}{2(\bar{A} + \delta \bar{B}) - p^2 \bar{D}} \qquad (58)$$

(+ for i = 1, - for i = 2). From the Y_{11_R}, the element values of the predistorted lossless circuit of Fig. 2 were obtained by ladder-expansion.

The lossy response of the network is shown in Fig. 3. The computed value of the passband ripple was 0.3009 db, i.e., for all practical purposes the design objective was met.

As a comparison, a uniformly predistorted network ($d = \frac{d_L + d_C}{2}$) was also synthesized, from the voltage ratio of Eq. (54). After adding the semiuniform losses to this network, a passband ripple of

0.3325 db was obtained. This result requires some additional explanation. Comparing the expression for a semiuniformly dissipative reactive branch-impedance Z_{su} with that for the same branch having uniform losses with the same average loss constant (Z_u), one gets

$$\frac{Z_{su} - Z_u}{Z_{su}} \cong \frac{\delta}{p} \quad . \tag{59}$$

Hence, if throughout the passband

$$\left|\frac{\delta}{p}\right| \leq \frac{1}{2Q_L} + \frac{1}{2Q_C} << 1, \tag{60}$$

uniform predistortion gives a very good approximation. Since for the networks treated in this paper the frequency-origin is not in the passband, the semiuniform predistortion procedure described in this paper is needed only if the Q's are very low or the required frequency-response is very sharp.

Acknowledgement

The writer is indebted to Mr. B.R. Smith and Mr. M. Stolarik for working out the example of Sec. V, and to Mr. M. Sablatash for useful discussions.

References

[1] S. Darlington, "Synthesis of reactance 4-poles which produce prescribed insertion loss characteristics", J. Math. and Phys., vol. 18, pp. 257-353; September, 1939.

[2] W. Cauer, "Theorie der linearen Wechselstromschaltungen", Akademie-Verlag, Berlin, Chapter XX (by Nai-Ta Ming), pp. 1081-1130; 1960.

[3] R. Saal and E. Ulbrich, "On the design of filters by synthesis", IRE Trans. on Circuit Theory, vol. CT-5, pp. 284-327; December, 1958.

Footnotes

1. Also, the orders of the corresponding lossy and lossless networks turn out to be different in all cases; this constitutes another contradiction.

2. This displacement of the pole at *p*=o results in a second-order small change of the passband attenuation response, as a simple derivation (using first-order Taylor expansions) easily shows.

3. This statement disregards poles occuring in origin-symmetric pairs, since such pairs contribute only even factors to the denominators of Λ and Λ_R.

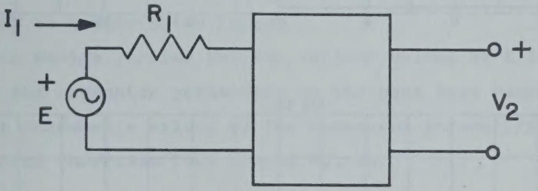


FIG. 1

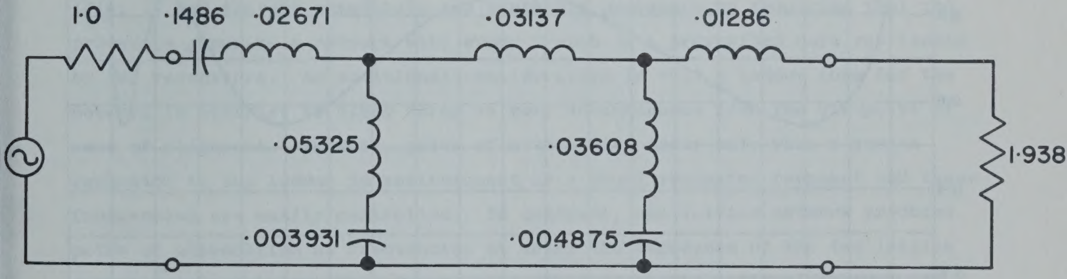

1·0
·1486
·02671
·03137
·01286
·05325
·03608
1·938
·003931
·004875

FIG.2

FIG 3a

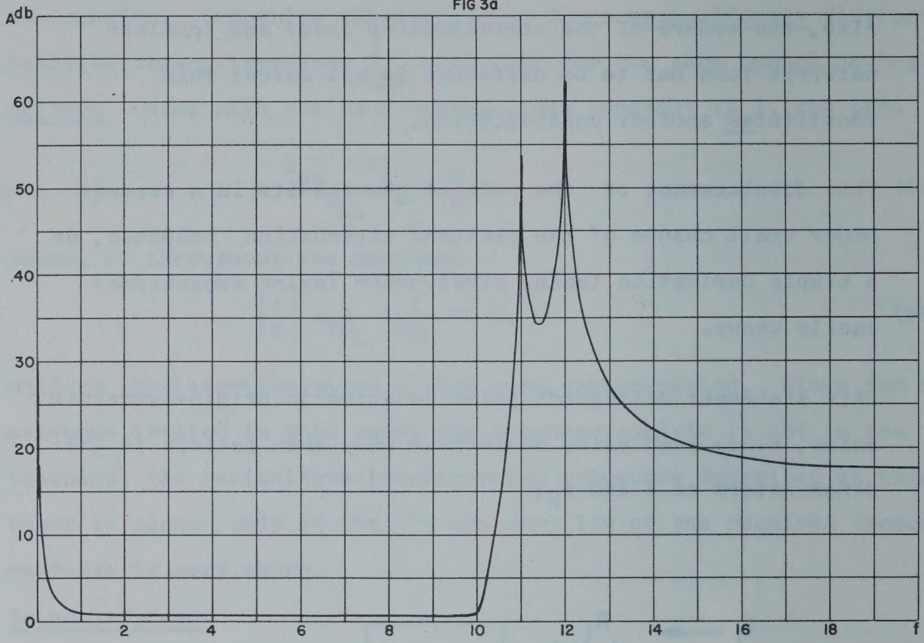
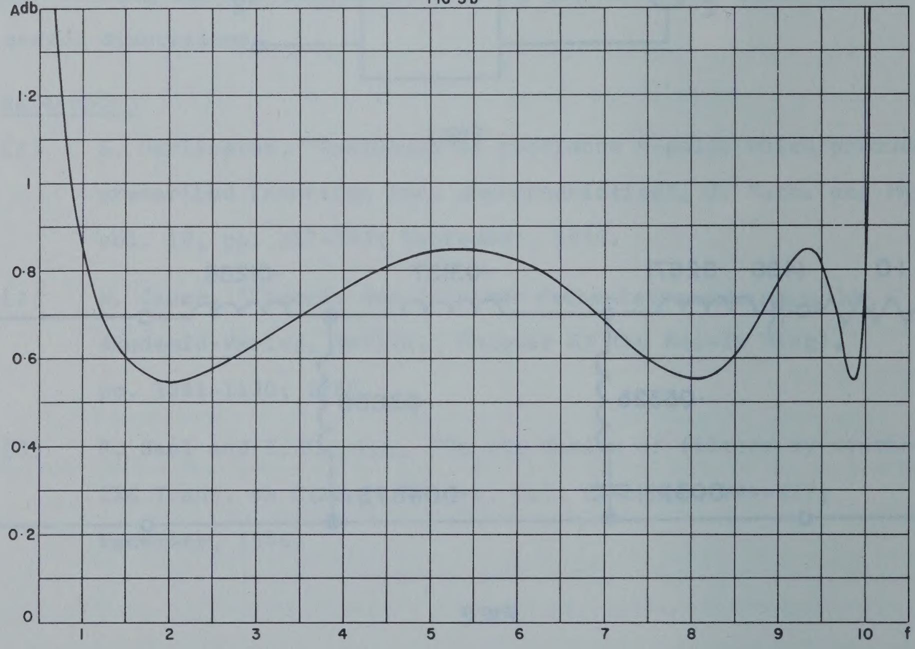

Adb
60
50
40
30
20
10
0
2
4
6
8
10
12
14
16
18
f

FIG 3b


Adb
1·2
1
0·8
0·6
0·4
0·2
0
1
2
3
4
5
6
7
8
9
10
f

ON THE EXISTENCE AND REALIZATION OF CRYSTAL LADDER FILTERS

JAMES D. SCHOEFFLER*
Case Institute of Technology
Cleveland Ohio

ABSTRACT

The high performance of piezoelectric resonators make them desirable circuit elements in filters. Because of the constrained form of those elements, design of such filters is not straightforward and has in the past been done by either the image method or through the use of the symmetrical lattice configuration, resulting in little control of the pass band shape or low stop band attenuation and high tolerance requirements.

In this paper, a technique for the realization of prescribed transmission functions is derived which results in a ladder network of resonators terminated at each end in resistance. The class of transmission functions which can be so realized is derived and shown to be of the form

$$|\tau|^2 = [1 + (KM/N)^2]^{-1}$$

where N and M are odd and even polynomials whose zeros satisfy certain simple constraints provided that the constant K, which determines the peak pass band insertion loss, is sufficiently large.

An example design carried out for various values of K illustrates the dependence of the resonator parameters on the peak pass band insertion loss and indicates that reasonable values of the resonator parameters result for values of peak pass band insertion loss around 0.1 db.

1. Introduction

The high electrical Q of piezoelectric circuit elements such as quartz or ceramic resonators make them desirable circuit elements for filtering purposes.[1,2,3] These elements, however, have an equivalent circuit of a fixed form (Fig. 1) and thereby complicate any synthesis procedure by requiring that the procedure generate a network with every branch in a prescribed form realizable by the resonators. An additional consideration is that a ladder form for the network in contrast to other forms is very advantageous from the viewpoint of ease of alignment. That is, poles of attenuation occur only when a series resonator in the ladder is antiresonant or a shunt resonator resonant and these frequencies are easily controlled. In contrast, the lattice network produces poles of attenuation at frequencies at which the impedance of the two lattice arms become equal and these frequencies are not usually critical frequencies of the individual resonators and are thus harder to control.

The major disadvantage of the ladder network is that it is not quite as general a network as is the lattice. That is, any symmetrical network realizable at all is realizable in the form of a symmetrical lattice without the use of transformers whereas the ladder network does not permit realization of all symmetrical networks unless transformers are allowed.[4] For the large class of insertion loss functions which are realizable by a ladder network, it is in general easier to build and align the filter in the ladder form as well as less costly. Thus it is worthwhile investigating the conditions under which a ladder network of resonators exists as a realization of an insertion loss function.

---

*Engineering Division, Case Institute of Technology, Cleveland, Ohio and Consultant, Clevite Corporation, Cleveland, Ohio.

In order to achieve maximum stop-band attenuation per resonator and maximum control over pass-band shape and skirt-ratio, it is necessary to know the conditions under which an insertion loss function is realizable by a ladder structure with resonator elements. This is even more essential if the design is to be automated on a digital computer. The purpose of this paper is to derive a synthesis scheme for crystal ladder filters together with conditions on the insertion loss function sufficient to guarantee realizability in the form of a ladder network of resonators.

II. Synthesis of the Ladder Network

The design of a filter from the insertion loss viewpoint starts with a specification of the transmission function $|\tau(\omega)|^2$. The insertion loss of the filter is defined to be 10 times the negative logarithm of the transmission. The necessary and sufficient condition for the existence of a realizable network with transmission function $|\tau(\omega)|^2$ is that the transmission be less than or equal to unity for all frequencies.[5]

Realization of the specified insertion loss in the form of a lossless network terminated at each end in a resistance is accomplished[6] by first finding the magnitude of the reflection coefficient $|\rho(j\omega)|^2$ and then $\rho(s)$ itself:

$$|\rho(j\omega)|^2 = 1 - |\tau(j\omega)|^2 \qquad 2.1$$

$$\rho(s)\,\rho(-s) = 1 - \tau(s)\,\tau(-s) \qquad 2.2$$

The factoring of the denominator of Eq. 2.2 is unique since $\rho(s)$ must have only left half s-plane poles. The numerator of $\rho(s)$ is not so restricted and hence the factoring is not unique. In general, different networks result from different factoring.

The driving point impedance of the lossless network terminated in a resistance is then given by[6]

$$Z(s) = \frac{1 - \rho(s)}{1 + \rho(s)} \qquad 2.3$$

If this impedance is to be realized in the form of a ladder network of resonators as in Fig. 2, the synthesis can proceed in the following manner.

Consider a cycle starting at some point ρ in the network (Fig. 2). The form of the network is clearly such that Z(s) must have a pole at the origin and Y(s) a pole at infinity and these poles can be used to advantage in the synthesis. The first step in the realization is the partial removal of the pole at the origin in the form of a series capacitor (Fig. 3A) in order to create a zero of impedance at one of the transmission zeros. It is shown in Appendix I that assuming the shunt resonators realize transmission below the pass band and series resonators the transmission zeros above the pass band, the value of the series capacitor removed is given by

$$1/C = \text{Minimum} \left\{-X_{-k}\omega_{-k}\right\} \qquad 2.4$$

where ω_{-k} is the kth transmission zero below the pass band, X_{-k} is the reactance of Z at this frequency (the real part of Z is zero at each transmission zero) and where the minimum operation is taken only over those transmission zeros at which the reactance is negative. Removing this capacitance insures realizability in the desired form if the original impedance was so realizable. It is shown in Appendix I that if Z(s) satisfies certain conditions, there is at least one transmission zero at which the reactance is negative and that the value of the capacitance removed is such that the residue of the pole of Z(s) at the origin is sufficiently large. Hence it is always possible to create a zero of impedance at one of the transmission zeros by partial removal of the pole at the origin.

The second step in the realization is to invert the remainder impedance and totally remove the pole of the admittance at the transmission zero determined from Eq. 2.4 in the form of a series resonant circuit (Fig. 3b). The next step consists of partially removing the pole of this admittance at infinity (since Z(s) is zero at infinity, the admittance has a pole there) and realizing it in the form of a shunt capacitor (Fig. 3c). This capacitance is chosen so that a zero of admittance is created at one of the remaining transmission zeros in the upper stop band. Realizability is assured if this capacitance is determined from

$$C = \text{Minimum} \left\{B_k/\omega_k\right\} \qquad 2.5$$

where ω_k is the kth transmission zero in the upper stop band and B_k is the susceptance of the admittance at this frequency and where the minimization operation is taken only over those frequencies at which the susceptance is positive. As before, at least one frequency exists at which the susceptance is positive and the value of C determined from Eq. 2.5 is less than the residue of the pole at infinity, assuming the conditions derived in the appendix are satisfied. Thus the realizability of the remainder is assured and the synthesis may continue.

Inverting the remainder admittance yields an impedance with a pole on the imaginary axis at a transmission zero and this can be removed as a parallel resonant circuit (Fig. 3d). At this point, the remainder impedance is of lower order and still satisfies the conditions of Appendix I and the cycle may be repeated. The initial and final capacitors in Fig. 2 which are not absorbed into resonators are necessary in order for the process to begin and end properly.

The synthesis procedure discussed here is very simple and straightforward to carry out and easily automated on a digital computer. The key to successful use of the technique lies in the specification of insertion loss functions which lead to impedances of the desired form. Existence conditions for general ladder networks have been studied elsewhere,[7,8] but these do not hold for the special case considered here. It is shown in Appendix I that for a ladder network with n resonators (n = 2 or more), the driving point impedance of the lossless network terminated in a resistance must have a pole at the origin, a zero at infinity, a denominator of order 2n + 2, an even part whose numerator has a second order zero at the origin, n second-order transmission zeros on the real frequency axis, and in addition, certain inequalities between the transmission zeros and the zeros of the even parts of the numerator and denominator of the impedance.

The synthesis procedure described above insures that at each step, the remainder impedance satisfies the conditions for realizability. Notice that the order in which the transmission zeros must be realized is determined by the realizability requirements. Notice also that these conditions are on the driving point impedance and not on the insertion loss function which must actually be specified. In the next section, these results are extended to the insertion loss function itself.

III. Conditions for the Realizability of the Insertion Loss Function

Of the four necessary and sufficient conditions listed in the appendix for the realizability of a driving point impedance as a ladder network of resonators, the first three to restrict the choice of the form of the insertion loss function and the fourth, the minimum pass band insertion loss which can be achieved for a given choice of transmission zeros and ones.

From the first condition, that Z(s) have a pole at the origin and a zero at infinity, it follows from Eq. 2.3 that the reflection factor, $\rho(s)$, must be +1 at $s=\infty$ and -1 at $s=0$.

The zeros of M(x) are the zeros of attenuation and can be of the form $(s^2+\beta^2)$ or $(s^4+\alpha s^2+\beta^2)$ or $(s^2-\beta^2)$ for real frequency zeros, complex zeros, or real-axis zeros respectively. Clearly $\rho(s)$ can have different signs at $s=0$ and $s=\infty$ only if M(s) has a factor of the form $(s^2-\beta^2)$ and also if there is an odd number of such factors. In general, it is desirable to choose the zeros of M(s) on the $j\omega$ axis and in the pass band, but one of them at least must be on the real-axis of the s-plane.

To achieve high stop band attenuation, low pass band attenuation, and small transition interval, an approximation similar to that used in Elliptic filters is desirable, namely that the transmission be of the form

$$\tau(s)\,\tau(-s) = \frac{1}{1 - [KM(s)/N(s)]^2} \qquad 3.1$$

where N(s) is an odd polynomial and M(s) is an even polynomial. Such a form insures that the transmission is less than unity at all real frequencies and hence is realizable in general. Assuming this form for the insertion loss, it follows that

$$\rho(s)\,\rho(-s) = \frac{K^2M^2(s)}{K^2M^2(s) - N^2(s)} \qquad 3.2$$

The second and third conditions for realizability specify the form and order of N(s) and M(s) to be:

$$KM(s) = K(s^2-\rho^2)\,(b_o+b_2s^2 + \ldots + b_{2n}s^{2n})$$

$$3.3$$

$$N(s) = +\, s(s^2+\omega_1^2)\,(s^2+\omega_2^2)\,(\ldots\ldots)\,(s^2+\omega_n^2)$$

where the ω_i are the real transmission zeros and K is a gain factor to be determined. The order of M(s) follows from the requirement that M/N have a pole at infinity since the transmission must be zero here.

The derivation of the realizability conditions assumed that the resonators in the series branches produce transmission zeros in the upper stop band and the ones in the shunt branches produce transmission zeros in the lower stop bands. Hence if n is even, the transmission zeros in Eq. 3.3 must be half above and half below the pass band and if n is odd, the number in each stop band must differ only by one.

The imposition of the fourth and last realizability condition onto the insertion loss function results in a limitation on the minimum achievable pass band insertion loss. In terms of the constant K in Eq. 3.3, the result is the

Theorem: An insertion loss of the form of Eq. 3.1 with M and N of the form of Eq. 3.3 is realizable in the form of a ladder network of resonators terminated at each end in resistance for K sufficiently large, that is, for sufficiently large pass band insertion loss. To prove the theorem, consider a factoring of Eq. 3.2 to determine $\rho(s)$:

$$\rho(s) = \frac{KM(s)}{R(s)} \qquad (3.4)$$

The polynomial R(s) contains all of the left-half plane zeros of $K^2M^2 - N^2$.

The impedance Z(s) is then given by

$$Z(s) = \frac{R-KM}{R+KM} = \frac{P(s)}{Q(s)} \qquad 3.5$$

Notice from Eq. 3.5 that the polynomials of interest, m_1 and m_2, are m_R-KM and m_R+KM respectively, where m_R is the even part of R(s). The conditions on m_1 and m_2 are most easily shown by first showing that the zeros of m_R lie in the pass band for K sufficiently large. For K sufficiently large, the zeros of $K^2M^2-N^2$ are arbitrarily close to those of M^2 as is well known from the usual root-locus arguments. Left half plane roots are those of R(s). Roots near the $j\omega$ axis roots of M(s) yield factors of the form

$$(s^2 + a_ks + b_k^2) \qquad 3.6$$

where b_k is in the pass band and a_k is arbitrarily small for K sufficiently large. The real s-plane axis roots of $M(s)^2$ (due to the factor $s^2-\beta^2$) contributes two real roots or a factor of the form

$$(s + A)(s + B) = s^2 + (A + B)s + AB \qquad 3.7$$

Assuming that β is in the pass band, the even part of the last factor has roots in the pass band for K sufficiently large.

If P and Q contain more than one factor, this result can be proved by induction. Suppose P or Q is the polynomial m + n and the zeros of m are in the pass band. Adding another term of the form of Eq. 3.6 yields the new polynomial.

$$m' + n' = (s^2 + as + b^2)\ (m + n) \qquad 3.8$$

and the new even part is m' given by

$$m' = (s^2 + b^2)\ m + asn \qquad 3.9$$

The zeros of m' are the roots of

$$\frac{m}{n} + \frac{as}{s^2 + b^2} = 0 \qquad 3.10$$

For K sufficiently large, the parameter "a" can be made as small as desired and hence the zeros of m' are arbitrarily close to the zeros of m together with a zero arbitrarily close to "b" which is in the pass band also. Hence if the zeros of m are in the pass band, so are the zeros of m' for K sufficiently large. Since the zeros of m were shown to be in the pass band for a second order polynomial, it follows that the zeros of m_R lie in the pass band for K sufficiently large. Since the degrees of the polynomials m_R and M are the same and since m_R is the even part of a Hurwitz polynomial, m_R is positive below the pass band, M is negative below the pass band (due to the factor $s^2 - \beta^2$) and both are the same sign above the pass band. Consequently, $m_1 = m_R - KM$ cannot have a zero below the pass band and $m_2 = m_R + KM$ cannot have a zero above the pass band since m_R has none there for K sufficiently large. Clearly then, choosing K large enough forces satisfaction for the inequality conditions on the zeros of m_1, m_2 and N.

This result does not give any information about the minimum pass band insertion loss that is realizable. In the specific case of elliptic filters (low pass), the minimum pass band insertion loss of mid-series low pass filters has been derived by Meinguet and Belevitch[4] but the result is not explicit (it is given as the solution to a transcendental equation) and moreover is not applicable to the general band pass resonator filter.

Computationally, the value of K is selected to give the desired pass and stop band attenuation limits. Then the resulting impedance is derived and checked for satisfaction of the inequalities. If the K selected was not large enough, the machine can select another K and repeat the calculation.

Since excessive attenuation in the pass band is not permitted and since it would not be desirable to achieve the lowest K since the element values would tend to be excessive, the choice for K is not critical and does not result in excessive computer time.

IV. Example and Conclusions:

As an example of this synthesis procedure, a number of three resonator filters were designed with a pass band and attenuation poles located as shown in Table 1, and insertion loss as shown in Fig. 4. Table 2 shows the parameters of the resulting filter for various values of the peak pass band insertion loss. This table indicates that quite reasonable resonator parameters result and that the parameters are not critically dependent upon the peak pass band insertion loss once a certain minimum has been passed. Computing time for the example was approximately 4 seconds per filter on a Univac 1107 machine.

The conditions under which an insertion loss may be realized by a ladder network of resonators terminated at each end in resistances are not excessively stringent but because of the requirement that one pair of transmission ones be on the real axis of the s-plane, the conventional approximations to ideal filter characteristics are not acceptable insertion loss functions. It is also evident that any insertion loss function for the band pass filter which is derived from a low pass function by the conventional low-pass to band-pass transformation is not an acceptable function. Consequently, the approximation problem must be investigated further to derive suitable insertion loss functions which satisfy the realizability condtions.

APPENDIX

In this appendix, the necessary and sufficient conditions for realization ofa driving point impedance in the form of a ladder network of resonators terminated in a resistance is derived. The conditions are summarized by the

Theorem: The positive real impedance $Z(s) = \frac{m_1+n_1}{m_2+n_2}$ is the driving point impedance of a ladder network of resonators terminated in a resistance (Fig. 2) with a band-pass characteristic and such that each resonator in a series arm has an anti-resonant frequency above the pass-band and each resonator in a shunt arm a resonant frequency in the lower stop-band if and only if

1. Z(s) has a pole at the origin and a zero at infinity
2. The polynomial m_2+n_2 is of order 2n+2 where n is the number of resonators
3. The numerator of the even part of Z is of the form

$$M_o(s) = - s^2(s^2 + \omega_1^2)^2 \ (......)(s^2 + \omega_n^2)^2$$

where all of the ω_i are real with half above and half below the pass band if n is even or with $(n\pm1)/2$ above and $(n\mp1)/2$ below if n is odd.

4. Arrange the n transmission zeros and the n non-zero zeros of the polynomials m_1 and m_2 ($\beta_1, \beta_2 \ldots, \beta_n$ and $\alpha_1, \ldots., \alpha_n$ respectively) in ascending order and number them from 1 to n.

Then if n is even, it is required that

$$\alpha_i \leq \omega_i \text{ for } i = (n/2) + 1, \ldots\ldots, n$$

$$\beta_i \geq \omega_i \text{ for } i = 1, 2, \ldots\ldots, n/2$$

and if n is odd,

$$\alpha_i \leq \omega_i \text{ for } i = (n\pm1)/2, \ldots\ldots, n$$

$$\beta_i \geq \omega_i \text{ for } i = 1, 2, \ldots\ldots, (n\mp1)/2$$

with the upper sign used if the odd transmission zero is above the pass band, and the lower sign if it is below the pass band. The proof of statements 1 through 3 of the theorem is quite simple, and follows directly from examination of the network in Fig. 1. The important part of the theorem is the set of inequalities in part 4. To prove the necessity of these conditions, an argument similar to one Fujisawa[5] used may be used. Consider the two resonator circuit in Fig. 7. By direct calculation, the input impedance is given by

$$Z(s) = \frac{(s^2+a)(s^2+D) + sC_1(s^2+b)(s^2+d)}{s^2C_1C_2(s^2+e)(s^2+b) + sC_2(s^2+e)(s^2+e) + sC_1(s^2+b)(s^2+d)}$$

In this case,

$$m_1 = (s^2+a)(s^2+d)$$

$$m_2 = s^2C_1C_2(s^2+e)(s^2+b)$$

Hence the zeros are

$$\alpha_1 = \sqrt{e}$$

$$\alpha_2 = \sqrt{b}$$

$$\beta_1 = \sqrt{d}$$

$$\beta_2 = \sqrt{a}$$

The transmission zeros are

$$\omega_1 = \sqrt{d}$$

$$\omega_2 = \sqrt{b}$$

Clearly

$$\alpha_2 = \omega_2$$

and

$$\beta_1 = \omega_1$$

Thus the necessity for n = 2 is proved and can be proved for arbitrary n by induction. Assume the theorem is true for Z(s) given by

$$Z(s) = (m_1+n_1)/(m_2+n_2)$$

and consider the addition of a resonator with impedance

$$Z_{res}(s) = \frac{s^2+a}{sC(s^2+b)}$$

Since Z(s) has a pole at the origin by condition 1 of the theorem, it is possible to write

$$m_2+n_2 = s(p_2+q_2)$$

where

$$m_2 = sp_2 \text{ and } n_2 = sq_2 \text{ .}$$

By direct calculation, the new m_1 and m_2 are given by

$$m_1' = m_1C(s^2+b) + q_2(s^2+a)$$

$$m'_2 = p_2sC(s^2+b).$$

An additional transmission zero at $\omega = \sqrt{b}$ has now been added. The zeros of m_2' are the zeros of m_2 together with an additional one at $\omega = /b$, the new transmission zero above center frequency. Clearly then the ith zero of m_2' is less than or equal to the ith transmission zero for all transmission zeros above center frequency. The zeros of m_1' are different from those of m_1 and are the zeros of

$$\frac{m_1}{n_2} + \frac{s^2+a}{sC(s^2+b)} = 0$$

Since m_1/n_1 is a physically realizable reactance function and since only the zeros of m_1 below center frequency are of interest in the theorem and because the impedance of the resonator is negative in the lower stop band (by assumption the resonant and antiresonant frequencies are above the lower stop band), the zeros of m_1' are larger in magnitude than are the zeros of m_1 (because the addition of a negative reactance to a reactance function shifts the zeros upward to higher frequencies). Hence if the conditions of the theorem applied to m_1, it must apply also to m_1'. That is, if the ith zero of m_1 larger than its corresponding transmission zero, so must the ith zero of m_1' which is larger than the corresponding zero of m_1.

In similar fashion it is easy to prove that the addition of a shunt resonator does not affect the zeros of the polynomial m_1 but causes all the zeros of the polynomial m_2 which are above the lower stop band to decrease in magnitude, and hence the condition 5 of the theorem still applies.

To prove sufficiency, the impedance must be synthesized. Assume Z(s) is given which satisfies the conditions of the theorem. It is first necessary to partially remove the pole at the origin to cause the remainder impedance to be zero at one of the transmission zeros in the lower stop band so that the synthesis can continue. It is important to realize the transmission zeros in the correct order for otherwise the conditions of the theorem may be invalidated at some later point in the design and the synthesis will fail.

Let the impedance be $Z(s)$ and consider the remainder impedance $Z_1(s)$ defined by

$$Z_1(s) = Z(s) - A/s$$

Since the pole is only partically removed, the order of the polynomials of Z are not changed and it is necessary only to insure that statement 4 of the theorem remains fulfilled. Note that the new polynomials m_1 and m_2 will be

$$m_1' = m_1 - Aq_2$$

$$m_2' = m_2$$

Clearly the zeros of m_2 are not affected. The zeros of m_1' are the zeros of

$$\frac{m_1}{n_2} - \frac{A}{s} = 0.$$

Subtracting A/s essentially adds a positive reactance to m_1/n_2, causing the zeros of m_1 to shift to lower frequencies. In order that the condition of statement 4 of the theorem remain fulfilled, A must be sufficiently small so that no zero of m_1 shifts beyond its corresponding transmission zero. If A is chosen to be

$$A = \text{Minimum} \left\{ - X_k \omega_k \right\}$$

where X_k is the reactance of Z at the kth transmission zero ω_k and the minimization operation is taken only over those frequencies in the lower stop band at which the reactance is negative, then no zero of m_1 can shift below its corresponding transmission zero.

It is necessary that the value for A so determined is not larger than the residue of the pole at the origin and that at least one transmission zero below center frequency exists at which the reactance is negative before the synthesis can be guaranteed to succeed. If the network is to contain resonators only as in Fig. 1, then none of the ω_i can be zero. Since the smallest transmission zero is below the lowest non-zero zero of m_1, and because m_1/n_2 is physically realizable, it follows that the reactance of m_1/n_2 at the lowest transmission zero is negative. Since at this transmission zero Z equals m_1/n_2, the reactance of Z is negative here and it has been shown that one transmission zero exists at which the reactance is negative.

Because Z has a pole at the origin, $z_{22} = m_2/n_2$ does not have a pole at the origin and hence the residue of m_1/n_2 is the residue of Z at the pole at the origin. As a consequence, the reactance of Z at ω_1 is less in magnitude than required so that the A removed is less than the residue of the pole and the remainder is guaranteed realizable.

After removing the series capacitor, the remainder impedance (which satisfies all the conditions of the theorem) is inverted and the pole on the $j\omega$ axis at the transmission zero determined from the minimization operation above may be totally removed to realize that transmission zero.

Total removal of this pole leaves the pole at the origin and the zero at infinity, and reduces the order of the denominator of Z by 2 as required. The zeros of the even part of Z are not affected except for the one which has been realized and this is removed. Hence conditions 1, 2, and 3 of the theorem remain fulfilled. To show that the last condition is still fulfilled, let $j\omega_o$ be the transmission zero being realized. Since it is a zero of m_1+n_1 because of the previous zero shifting step, the remainder admittance becomes

$$Y - \frac{ks}{s^2+\omega_o^2}$$

The even part of the denominator is $m_1' = m_1/(s^2+\omega_o^2)$. Clearly the zeros of m_1' are identical to those of m_1 except for the zero just realized and hence if condition 4 was previously satisfied, it still must be. The even part of the numerator is m_2' and is given from the previous equation by

$$(s^2+\omega_o^2)\ m_2' = m_2 - ksn_1/(s_2+\omega_o^2)$$

The zeros of m_2' are the zeros of

$$\frac{m_2}{m_1} - \frac{ks}{s^2+\omega_o^2} = 0$$

Above the pass band, the term $-ks/(s^2 + \omega_o^2)$ is always positive reactance and hence all the zeros of m_2' are lower in frequency than those of m_2 and therefore if the condition 4 was previously satisfied, it still is satisfied. Thus the remainder impedance also satisfies the conditions of the theorem and the realization may continue. Then the shunt capacitor of that resonator is realized by partially removing the pole at infinity. The amount of the pole at infinity that is removed is determined by the condition that the remaining admittance must be zero at one of the transmission zeros in the upper stop band so that the synthesis can continue. The size of C to be removed is given by

$$C = \text{Minimum } B_k/\omega_k$$

where ω_k is the kth transmission zero in the upper stop band and B_k is the susceptance of Y at that frequency and where the minimization operation is taken only over those ω_k where B_k is positive. This choice for C is again to insure that no zero of m_2 shall shift upward beyond its corresponding transmission zero, thereby invalidating condition 5 of the theorem. That the value of C so determined exists and is less than the residue of the pole at infinity is proved in the same way as previously and is not repeated here.

It should be noted that if transmission zeros at zero or infinite frequency were allowed, the value of the series or shunt capacitance determined from the minimization operation may exceed the residue of the pole at the origin or infinity and in that case the pole would be totally removed and the resulting realization would not be completely resonator-realizable. This is the reason all transmission zeros must be finite non-zero.

By showing that the network can be synthesized, the sufficiency of the theorem is shown and the proof is complete.

It should be noted that the conditions under which an impedance is realizable by a network of resonators is more stringent than the conditions derived by Fujisawa. He showed that for low pass filters of the mid-series or mid-shunt variety, inequality constraints between the zeros of n_1 or n_2 and the transmission zeros exists similar to those in statement 5 of the theorem here. If his low pass filters are transformed to band pass filters by means of a low-pass to band-pass transformation, the conditions become simply that all of the zeros of m_1 (or m_2) must be larger than their corresponding lower stop band transmission zero whereas in the resonator case, some of the zeros of both m_1 and m_2 must fulfill inequality constraints. This is to be expected since the resonator ladder circuit is a special case of a band-pass ladder network.

REFERENCES

1. T.R. O'Meara, "The Synthesis of a Class of Wide-Band Lattice Crystal Filters with Symmetrical Insertion Loss," I.R.E. Trans., Vol. CT-7, June, 1960, pp. 121-127.

2. J. Schoeffler, D. Curran, W. Szczecinski, "The Design of Piezoelectric Ceramic Resonator Filters with Prescribed Insertion Loss," Proc. Nat'l. Components Conf., 1961.

3. J. Schoeffler, "Insertion Loss Design of Symmetrical Lattice Piezoelectric Resonator Filters with Symmetrical or Unsymmetrical Pass Bands," Trans. I.R.E., Vol. CT-9, No. 3, Sept., 1962, pp. 251-256.

4. E.A. Guillemin, "Synthesis of Passive Networks," John Wiley and Sons, New York City, 1957, p. 39.

5. Guillemin, op. cit., p. 460.

6. Guillemin, op. cit., p. 458-467.

7. J. Mainguet, V. Belevitch, "On the Realizability of Ladder Filters," Trans. I.R.E., Vol. CT-5, No. 4, Dec., 1958, pp. 253 to 255.

8. T. Fujisawa, "Realizability Theorem for Mid-Series or Mid-Shunt Low-Pass Ladders without Mutual Induction," Trans. I.R.E., Vol. CT-2, pp. 320-324, Dec., 1955.

TABLE I

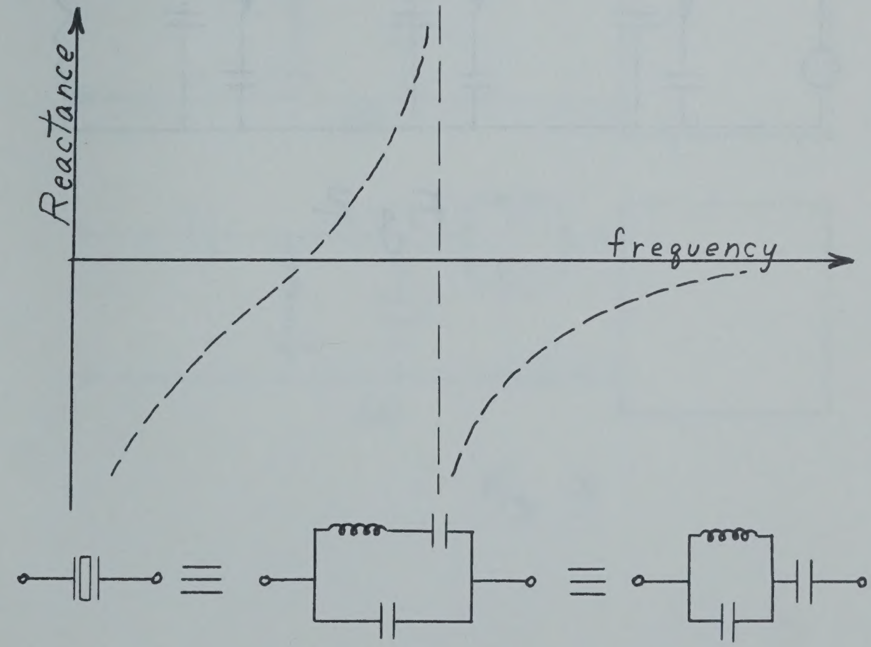
Specifications of Example Design

Bandwidth	5%
Center frequency	455 KC
Pass Band	443.9 to 466.1 KC
Source Resistance	1000 ohms
Number Resonators	3
Attenuation Poles	432.8, 437.2, 477.2 KC

TABLE II

Filter Parameters of Designs

Peak Pass Band Insertion Loss (db)	0.1	0.5	1.0	2.0
Terminal Resistance (ohms)	980	980	990	1000
$\Delta f/f_r$ Ratios (percent)	5.5	4.7	4.4	4.1
	3.5	4.2	4.4	4.5
	7.3	6.3	5.9	5.6
Resonator Capacities (pico farads)	248	405	496	653
	180	158	158	158
	360	541	676	879
Input Series Capacity (pico farads)	3357	1014	788	631
Output Series Capacity (pico farads)	1239	743	631	541



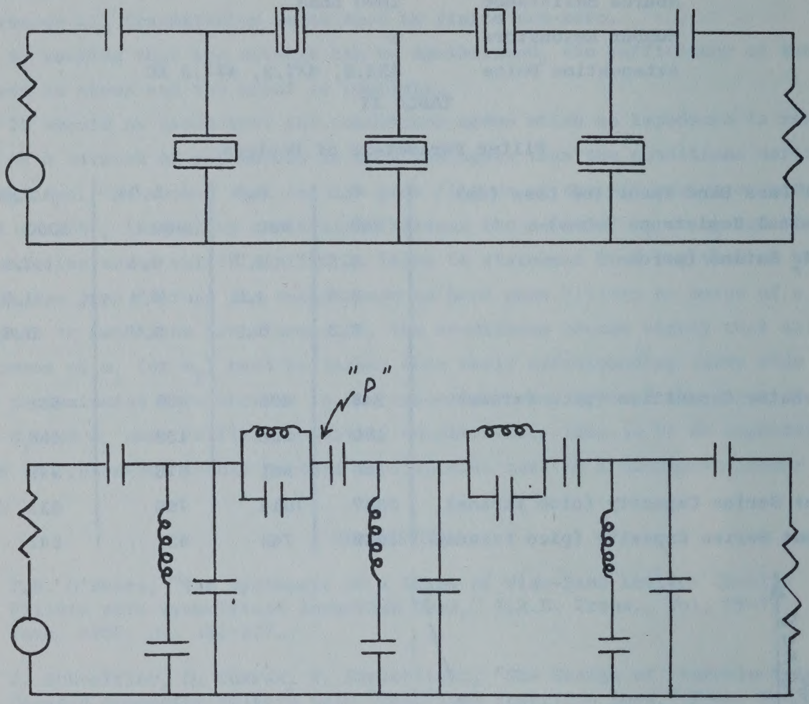


Fig 2

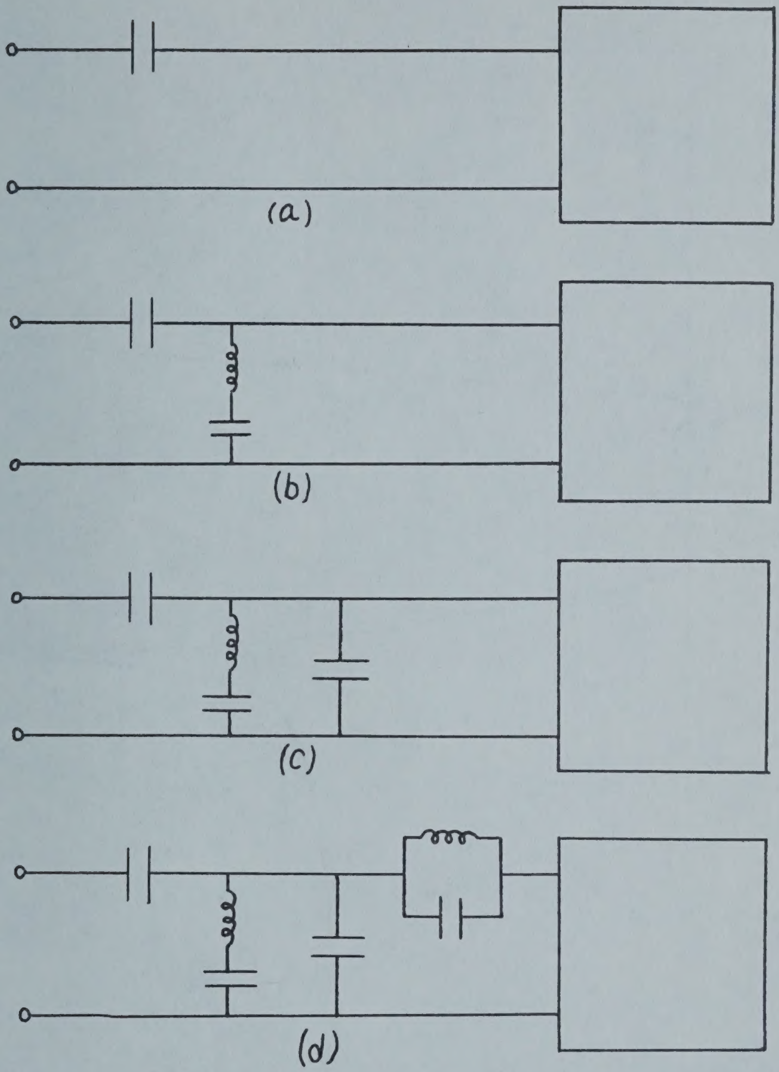


Fig. 3

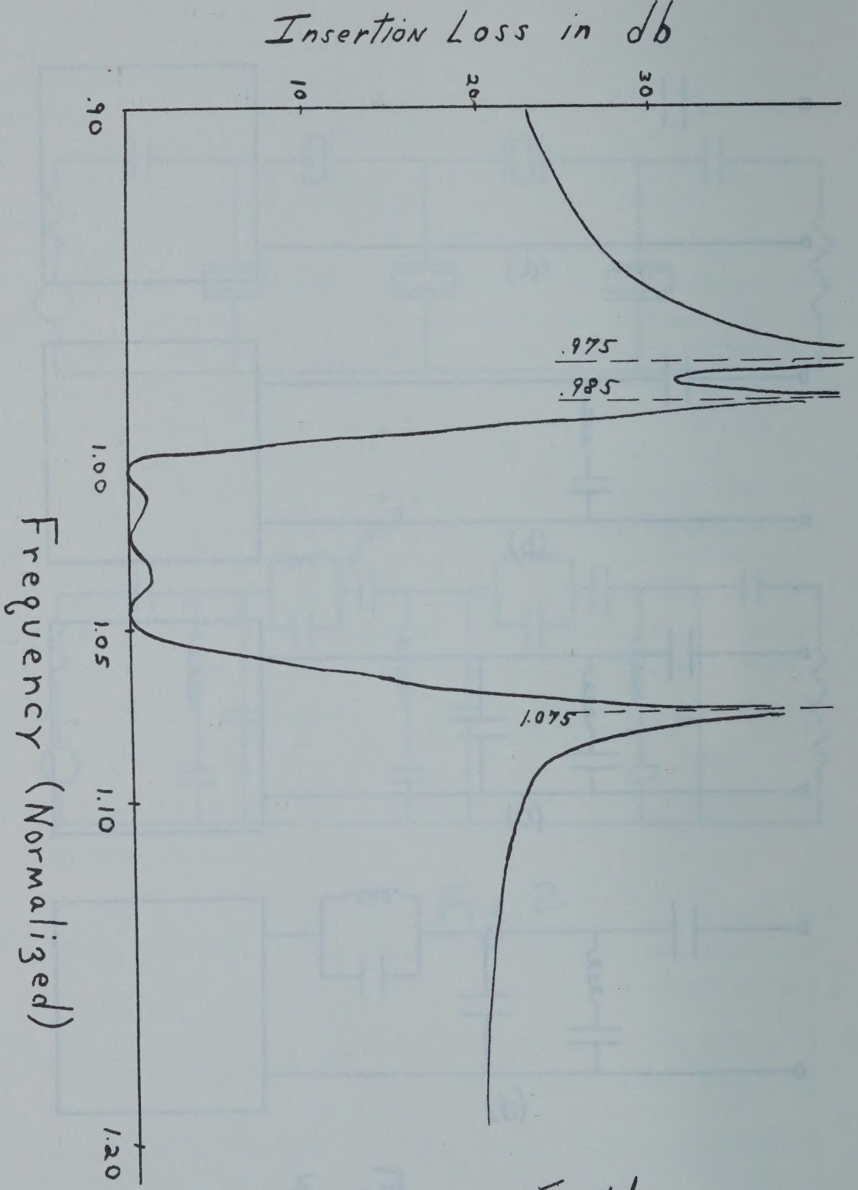

Insertion Loss in db
10
20
30
.90
1.00
1.05
1.10
1.20
.975
.985
1.075
Frequency (Normalized)

Fig 4

A NEW APPROACH IN TRANSFER-FUNCTION SYNTHESIS

WAYNE H. CHEN
University of Florida
Gainesville, Florida

Abstract.- This paper describes a new approach in transfer-function synthesis in which a ladder network configuration is first predicted and realized into a known network merely with the aid of simple, routine procedures. Minimum-phase transfer functions with purely imaginary or real-and-negative transmission zeros are realized with LC, RC, RL and RLC networks.

I. INTRODUCTION

The problem of synthesizing a given minimum-phase transfer function $Z_T(s) = E_2/I_1$ or $A(s) = I_2/I_1$ with the network arrangement in Fig. 1(a), or $Y_T(s) = I_2/E_1$ or $G(s) = E_2/E_1$, in Fig. 1(b), within a constant multiplier, is one of considerable interest to network as well as control system designers. Since $A(s) = -(E_2/R_2)/I_1 = -Z_T(s)/R_2$ and $G(s) = -I_2R_2/E_1 = -Y_T(s)R_2$, it suffices to consider only $Z_T(s)$ and $Y_T(s)$.

The reader's familiarity with the realizability conditions [1], [2], with ladders, of the open- and short-circuit parameters z_{ij} and y_{ij}, and the transfer functions is assumed.

A. A Preliminary Topic: Obtaining LC Ladder Parameters

With the assumed forms of $Z_T(s)$ and $Y_T(s)$ as ratios of two polynomials in s with real coefficients, the following relations

$$Z_T(s) = H\frac{P(s)}{Q(s)} = \frac{z_{21}}{1 + z_{22}}, \quad Y_T(s) = -H\frac{P(s)}{Q(s)} = \frac{y_{21}}{1 + y_{22}} \quad (1)$$

may be readily obtained for the normalized network arrangements with $R_2 = 1$ in Fig. 1. Based upon the realizability conditions of the parameters for LC ladders, the procedure for obtaining the parameters z_{21} and z_{22} from a given $Z_T(s)$ or y_{21} and y_{22} from a given $Y_T(s)$ may be given as follows:

1) Divide both the numerator $\pm HP(s)$ and denominator $Q(s)$ in (1) with the even (odd) part of $Q(s)$ if $P(s)$ is an odd (even) polynomial in s. 2) Identify them with the right-hand sides of (1), obtaining the expressions for z_{21} and z_{22} or y_{21} and y_{22} .

B. Another Preliminary Topic: Obtaining RC, RL or RLC Network Parameters

To synthesize a given transfer function with an RC, RL or RLC network arrangement as depicted in Fig. 1(a) or (b), we may first synthesize it in the form of Fig. 2(a) or (b), and then pull out R_2 to obtain the network arrangement of Fig. 1(a) or (b).

For the purpose of synthesis, we may use

$$z_{21} = H \frac{P(s)}{Q(s)}, \quad z_{22} = \frac{R(s)}{Q(s)} = \text{a chosen function} \qquad (2)$$

as the "associated parameters" of a given $Z_T(s) = H\, P(s)/Q(s)$ and realize them with the preliminary arrangement in Fig. 2(a), or

$$y_{21} = -H \frac{P(s)}{Q(s)}, \quad y_{22} = \frac{R(s)}{Q(s)} = \text{a chosen function} \qquad (3)$$

of a given $Y_T(s) = -H\, P(s)/Q(s)$ and realize, in Fig. 2(b).

The parameter z_{22} in (2) (or y_{22} in (3)) must be appropriately chosen as a realizable RC, RL or RLC impedance (or admittance) for RC, RL or RLC synthesis.

C. Description of Approach

The method of synthesizing a transfer function $Z_T(s)$ or $Y_T(s)$, or its associated parameters z_{21} and z_{22} (or z_{12} and z_{11}) or y_{21} and y_{22} with LC ladders was first introduced by

Cauer[3] and extended to RC ladders by Guillemin[4]. These methods are discussed in most of the recent synthesis books [2], and are generally referred as the zero-shifting methods.

A new approach of ladder synthesis is now described with the following steps: 1) Determine z_{21} and z_{22} or y_{21} and y_{22} from the given transfer function. 2) Predict one (or several) network configuration consisting of a number of transmission-zero producing basic sections. 3) Synthesize z_{22} or y_{22} as a driving-point function with the "predicted" network configuration, section by section, according to an established procedure for each section.

This "prediction" approach yields the same results as the zero-shifting methods, but in a more straightforward fashion. It may also be readily extended [5] for synthesizing transfer functions with complex transmission zeros and is a vehicle to accommodate different synthesis methods, e.g., Dasher's method [6], etc., in the solution of a single synthesis problem on an organized basis.

II. TRANSMISSION ZEROS

The zeros of a transfer function of the response/excitation type, e.g., $Z_T(s)$ or $Y_T(s)$, are called "transmission zeros." Assuming that z_{21} and z_{22} (or y_{21} and y_{22}) have identical denominators and the poles of z_{22} (or y_{22}) are also the poles of z_{21} (or y_{21}), (1) implies that the transmission zeros are the zeros of z_{21} or y_{21} .

A. Standard Network Configurations

For a prescribed set of parameters z_{21} and z_{22} (or y_{21} and y_{22}), we shall attempt to synthesize it with the standard network configuration having parallel (or series) branches at both ends as depicted in Fig. 3(a) (or 3(b)). The absence of series (or parallel) branches at both ends is assumed, for simplicity in results, and for the practical reasons to be given later in Section IV-B.

B. Transmission Zeros as Contributed by Circuit Branches

By obtaining the z_{21} (or y_{21}) parameter of the network configuration in Fig. 3(a) (or 3(b)), examining its poles, and verifying later with numerical illustrations in Section V, we may readily establish the following properties: 1) The poles of a series branch Z_k or a parallel branch Y_i in Fig. 3(a) or (b) are transmission zeros; these transmission zeros are said to be "contributed" or "produced" by Z_k or Y_i. 2) In general, a finite non-zero-value transmission zero, e.g., $s = -3$, is contributed by a series Z_k or a parallel Y_i, while a simple transmission zero at $s = 0$ or $s = \infty$ may be contributed by a number of branches, e.g., Y_1, Y_3, Y_5 or Y_1, Y_3, Z_4, Y_5, Z_6 in Fig. 3(a).

III. BASIC SECTIONS

Two prototype basic sections are introduced in Fig. 4 with a shaded box as its "principal branch" producing one or a pair of transmission zeros (in general, at $s \neq 0$ and $s \neq \infty$). The "auxiliary branch", depicted by a clear box in the figure, contributes to a transmission zero at $s = \infty$ or $s = 0$, or does not contribute at all.

A number of basic sections belonging to these prototypes are introduced in the following paragraphs.

A. Type 1 LC Section

Fig. 5(a) depicts a type 1 LC section whose principal branch $Y_a^{(k)}$ is responsible for a pair of purely imaginary transmission zeros, $s = \pm j\omega_k$, and whose auxiliary branch $Z_b^{(k)}$ contributes to a transmission zero at $s = \infty$ or $s = 0$. The super- or subscript k indicates that this type 1 LC section is used as the kth transmission-zero-producing section of a network.

The procedure of synthesizing a given Z_k with a type 1 LC section as depicted in Fig. 5(a), is given as follows:

For $Z_k(j\omega_k) = jX_k$, we have

$$L_b^{(k)} = \frac{X_k}{\omega_k} \qquad (X_k \geqslant 0), \qquad (4)$$

$$C_b^{(k)} = \frac{1}{\omega_k |X_k|} \qquad (X_k < 0), \qquad (5)$$

$$Z_b^{(k)} = sL_b^{(k)} \quad \text{or} \quad Z_b^{(k)} = \frac{1}{sC_b^{(k)}}, \qquad (6)$$

and

$$L_a^{(k)} = \frac{1}{M_k}, \qquad C_a^{(k)} = \frac{M_k}{\omega_k^2}, \qquad (7)$$

where

$$M_k = \left[Y_k^* \frac{(s^2+\omega_k^2)}{s} \right]_{s=j\omega_k}, \quad Y_k^* = \frac{1}{Z_k - Z_b^{(k)}} \qquad (8)$$

thus leaving

$$Z_{k+1} = \frac{1}{Y_k^* - Y_a^{(k)}} \qquad (9)$$

to be synthesized for the next section. This procedure is summarized in Table I for later use.

The above procedure may be readily justified in this manner: With $L_a^{(k)}$ and $C_a^{(k)}$ resonant at the transmission zero $s = j\omega_k$, i.e., a short circuit across 3 and 3' in Fig. 5(a), $[Z_k]_{s=j\omega_k} = [Z_b^{(k)}]_{s=j\omega_k}$ or $jX_k = Z_b^{(k)}(j\omega_k)$, which implies (4) for $X_k \geqslant 0$ or (5) for $X_k < 0$. To derive (7) and (8) is merely to obtain and identify a partial fraction $Y_a^{(k)} = M_k s/(s^2+\omega_k^2)$ from Y_k^* .

A type 1 LC section has two special cases in the form of Fig. 5(a) with the deletion of $C_a^{(k)}$ or $L_a^{(k)}$, whose principal branch $Y_a^{(k)} = 1/sL_a^{(k)}$ or $Y_a^{(k)} = sC_a^{(k)}$ produces a transmission zero at $s = 0$ or $s = \infty$ (instead of $s = \pm j\omega_k$).

B. Other Basic Sections

Fig. 5(b) depicts a type 2 LC section whose principal branch is responsible for a pair of purely imaginary transmission zeros, $s = \pm j\omega_k$, and whose auxiliary branch contributes to a transmission zero at $s = \infty$ or $s = 0$. Figs. 6 and 7 depict type 1, 2 RC and RL sections, whose principal branch is responsible a real negative transmission zero, $s = -\sigma_k$, and whose auxiliary (resistive) branch does not contribute to any transmission zero. The synthesis procedures of these basic sections are tabulated in Table I.

Like the type 1 LC section, the special cases of these basic sections have principal branches producing a transmission zero at $s = \infty$ or $s = 0$ (instead of $s = \pm j\omega_k$ or $s = -\sigma_k$).

IV. PREDICTION

A. Method of Prediction

For a given distribution of transmission zeros, we predict the network with basic sections and additional circuit branches in the following manner: 1) Fit it into one of the standard forms in Fig. 3. 2) Use the principal branches to produce the finite non-zero-value transmission zeros. 3) Use the auxiliary and/or additional branches to contribute to the transmission zero at $s = 0$ or $s = \infty$, or to make no contribution if there is no transmission zero at $s = 0$ or $s = \infty$.

Some slight modification of the above procedure is necessary if the distribution has no finite non-zero-value transmission zeros. In this instance, the special cases of type 1 and 2 sections, whose principal branches also contribute to the transmission zero at $s = 0$ or $s = \infty$, are used.

For a given distribution of transmission zeros, we may predict a number of configurations. It should be recognized that if we try to synthesize with all these configurations, some will result in negative elements as symptoms of non-realizability.

B. Desirable Configuration for RC, RL or RLC Networks

For a prescribed set of parameters z_{21} and z_{22} (or y_{21} and y_{22}), it is often desirable to use a configuration with parallel (or series) resistive branches at both ends for assimilating an internal admittance (or impedance) of the source and providing a resistive load (Section I-B). For simple orderly development, such a configuration may consist exclusively of type 2 (or type 1) RC or RL sections and a parallel (or series) resistive branch on the extreme left, as will be illustrated later in numerical illustrations in Figs. 9 and 10.

It is, however, taken for granted that we may at times need to rearrange the order of the transmission zeros to be realized with our predicted configuration.

V. SYNTHESIS PROBLEMS

A. Illustration of An LC Synthesis Problem

For

$$Y_T(s) = -H\frac{(s^2+1)(s^2+4)}{39s^5+33s^4+65s^3+51s^2+20s+12}, \tag{10}$$

we obtain with the aid of (1):

$$y_{21} = -\frac{H(s^2+1)(s^2+4)}{39s^5+65s^3+20s},\ y_{22} = \frac{33s^4+51s^2+12}{39s^5+65s^3+20s}. \tag{11}$$

With the transmission zeros, $s = \pm j1$, $s = \pm j2$ and $s = \infty$, we predict the configuration in Figs. 8(a). Letting $Z_1 = 1/y_{22}$ and applying the type 1 LC section procedure in (4) through (9) twice, we obtain the network arrangement for the given transfer function in Fig. 8(b).

Instead of the above prediction with two type 1 LC sections followed by a series inductance $L_b^{(3)}$ as in Fig. 8(a), we may also predict and realize another configuration with a series inductance ($L_b^{(o)} = 13/11$) followed by two type 2 LC sections ($C_a^{(1)} = 11/2$, $L_b^{(1)} = 12/55$, $C_b^{(1)} = 55/12$; $C_a^{(2)} = 5/4$, $L_b^{(2)} = 4/15$, $C_b^{(2)} = 15/16$).

B. Illustration of An RC Synthesis Problem

For

$$Z_T(s) = H \frac{(2s + 1)(s + 2)}{(s + 1)(2s + 5)} , \tag{12}$$

we obtain with the aid of (2):

$$z_{21} = \frac{H(2s + 1)(s + 2)}{(s + 1)(2s + 5)} , \quad z_{22} = \frac{1}{2} \frac{(2s + 3)(s + 3)}{(s + 1)(2s + 5)} \tag{13}$$

with z_{22} chosen for RC synthesis (Section I-B). With transmission zeros, $s = -1/2$ and $s = -2$, we predict the configuration in Fig. 9(a). Letting $Y_1 = 1/z_{22}$, applying the type 2 RC section procedure (Table I) twice, and pull out R_2 as the load (Section I-B), we obtain the network arrangement for the given transfer function in Fig. 9(b).

For real negative transmission zeros, i.e., the zeros of $Z_T(s)$, we may realize it with an RC ladder as shown above; an RL ladder if z_{22} is appropriately chosen as a realizable RL driving-point impedance; or possibly an RLC ladder consisting of type 1, 2 RC and RL sections, as illustrated below.

C. Illustration of An RLC Synthesis Problem

For

$$Z_T(s) = H \frac{(s + 3)(s + 9)}{s^2 + 10s + 27} , \tag{14}$$

we obtain with the aid of (2):

$$z_{21} = H \frac{(s + 3)(s + 9)}{s^2 + 10s + 27} , \quad z_{22} = \frac{s^2 + 8s + 18}{s^2 + 10s + 27} \tag{15}$$

where z_{22} is merely a realizable driving-point impedance. With transmission zeros, $s = -9$ and $s = -3$, we predict the configuration in Fig. 10(a) with two type 2 sections, and take RC or RL sections as determined by the synthesis procedures. Finding $Y_1 = 1/z_{22}$ non-realizable with a type 2 RC section (Table I) for the transmission zero $s = -\sigma_1 = -9$, we realize it with a type 2 RL section (Table I) leaving Y_2 to be realized. We then find Y_2 realizable with a type 2 RL section (Table I) for the transmission zero $s = -\sigma_2 = -3$, leaving $Y_3 = G_a^{(3)} = 4/5$ which is a resistive branch of 5/4 ohms. Pulling out $R_2 = 1/G_a^{(1)} = 3/2$ ohms as the load (Section I-B), we obtain the network arrangement for the given transfer function in Fig. 10(b).

The straightforwardness of the synthesis methods as demonstrated in the preceding illustrations also suggests the possibility that the "prediction" approach may provide a basis for computer design. For example, the computer may be programmed to try out 1) different orders of the transmission zeros to be realized with a predicted configuration (Section IV-B), or 2) even different possible predictions, in the synthesis of a given transfer function.

BIBLIOGRAPHY

[1] A. D. Fialkow and I. Gerst, "The transfer function of general two terminal-pair RC networks," Quart. Appl. Math., vol. 10, pp. 113-127; July, 1952.

[2] M. E. Van Valkenburg, "Introduction to Modern Network Synthesis," John Wiley and Sons, Inc., New York, N.Y.; 1960.

[3] W. Cauer, "Filters open circuited on the output side," Elek. Nachr.-Tech., vol. 16, no. 6, pp. 161-163; June, 1939.

[4] E. A. Guillemin, "A note on the ladder development of RC networks," Proc. IRE, vol. 40, pp. 482-485; April, 1952.

[5] W. H. Chen, "Synthesis of minimum-phase transfer functions having complex transmission zeros with ladder networks," Proceedings of the 1963 Allerton Conference on Circuit and System Theory (November 15-17, University of Illinois).

[6] B. J. Dasher, "Synthesis of RC transfer function as unbalanced two terminal-pair networks," IRE Trans. on Circuit Theory, vol, CT-1, pp. 20-34; December, 1952.

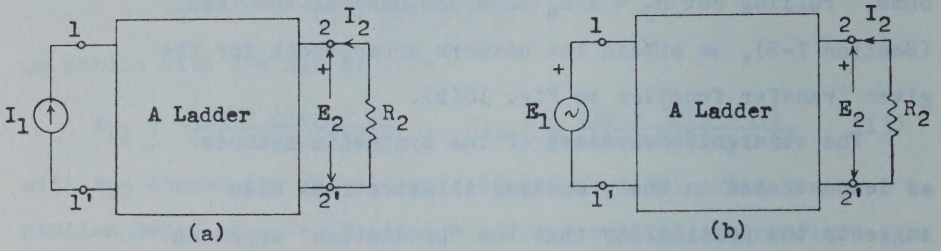

1
1'
I_1
A Ladder
2
2'
I_2
+
E_2
R_2
(a)
1
1'
+
E_1
A Ladder
2
2'
I_2
+
E_2
R_2
(b)

Fig. 1—Network arrangements for: (a) prescribed $Z_T(s)$ or z_{21} and z_{22} . (b) prescribed $Y_T(s)$ or y_{21} and y_{22} .

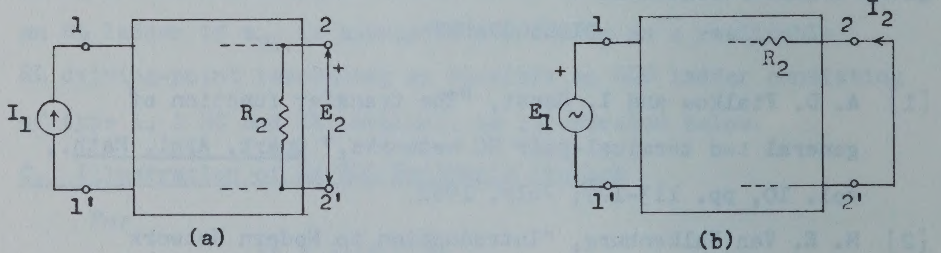

1
1'
I_1
R_2
2
2'
+
E_2
(a)
1
1'
+
E_1
R_2
2
2'
I_2
(b)

Fig. 2—Preliminary arrangements for: (a) prescribed $Z_T(s)$ or z_{21} and z_{22} . (b) prescribed $Y_T(s)$ or y_{21} and y_{22} .

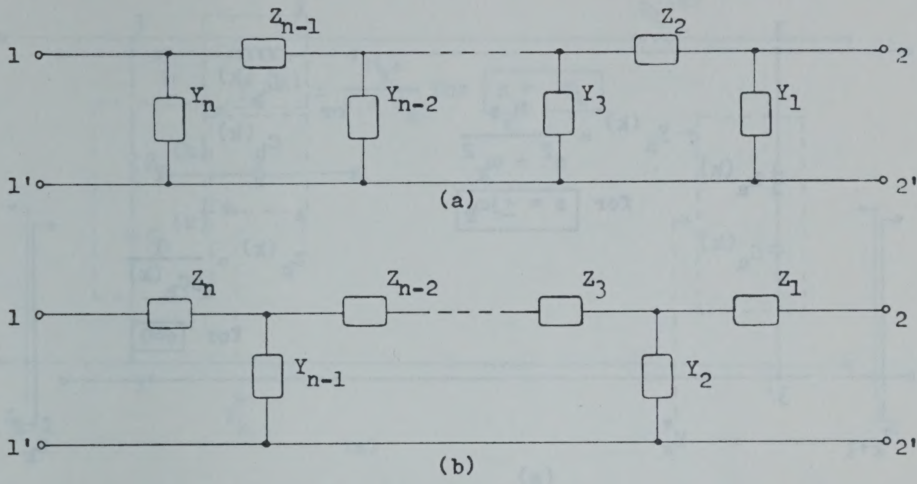

Z_{n-1}
Z_2
1
2
Y_n
Y_{n-2}
Y_3
Y_1
1'
2'
(a)
Z_n
Z_{n-2}
Z_3
Z_1
1
2
Y_{n-1}
Y_2
1'
2'
(b)

Fig. 3—Standard network configurations for: (a) prescribed z_{21} and z_{22} . (b) prescribed y_{21} and y_{22} .

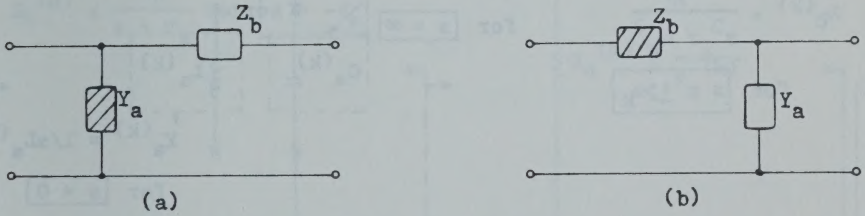

Z_b
Y_a
(a)
Z_b
Y_a
(b)

Fig. 4—Prototype basic sections with shaded boxes depicting principal branches and clear boxes, auxiliary branches. (a) Type 1 section. (b) Type 2 section.

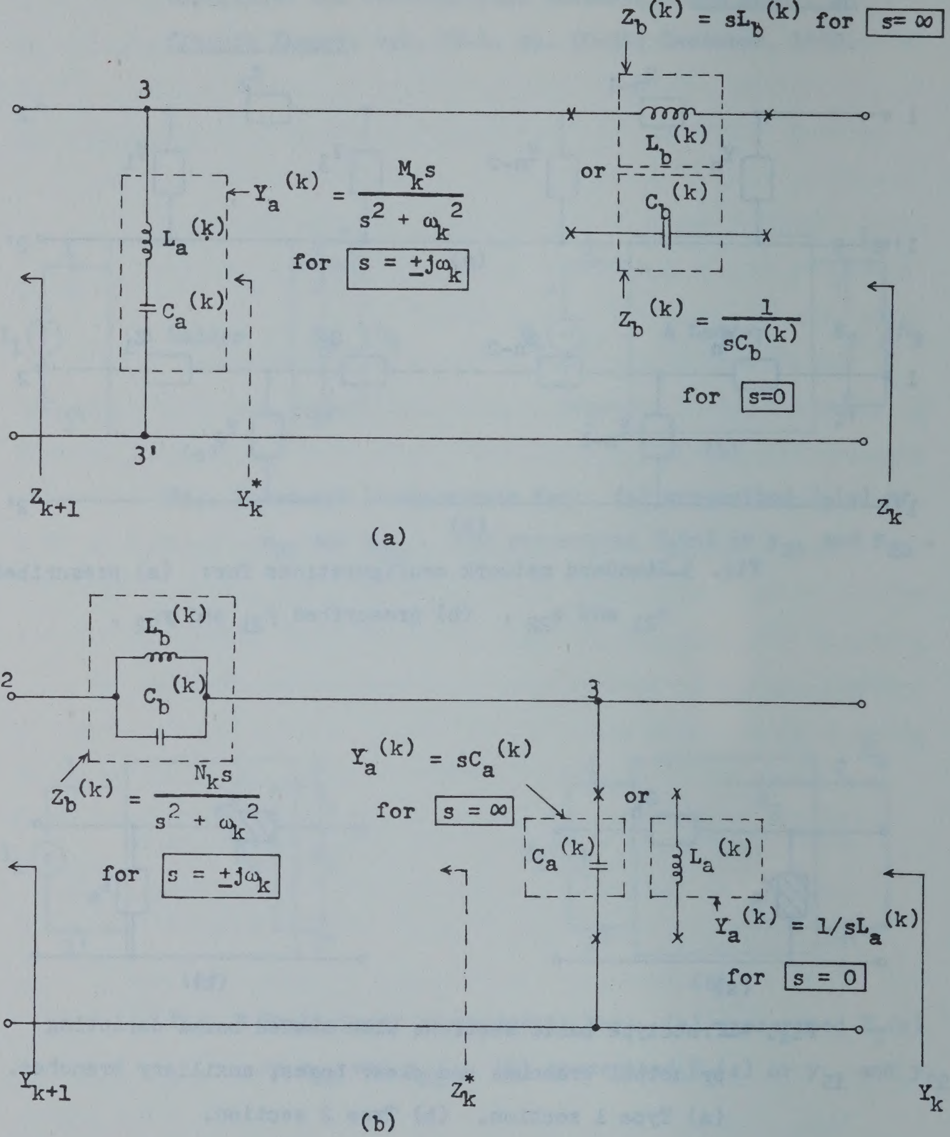
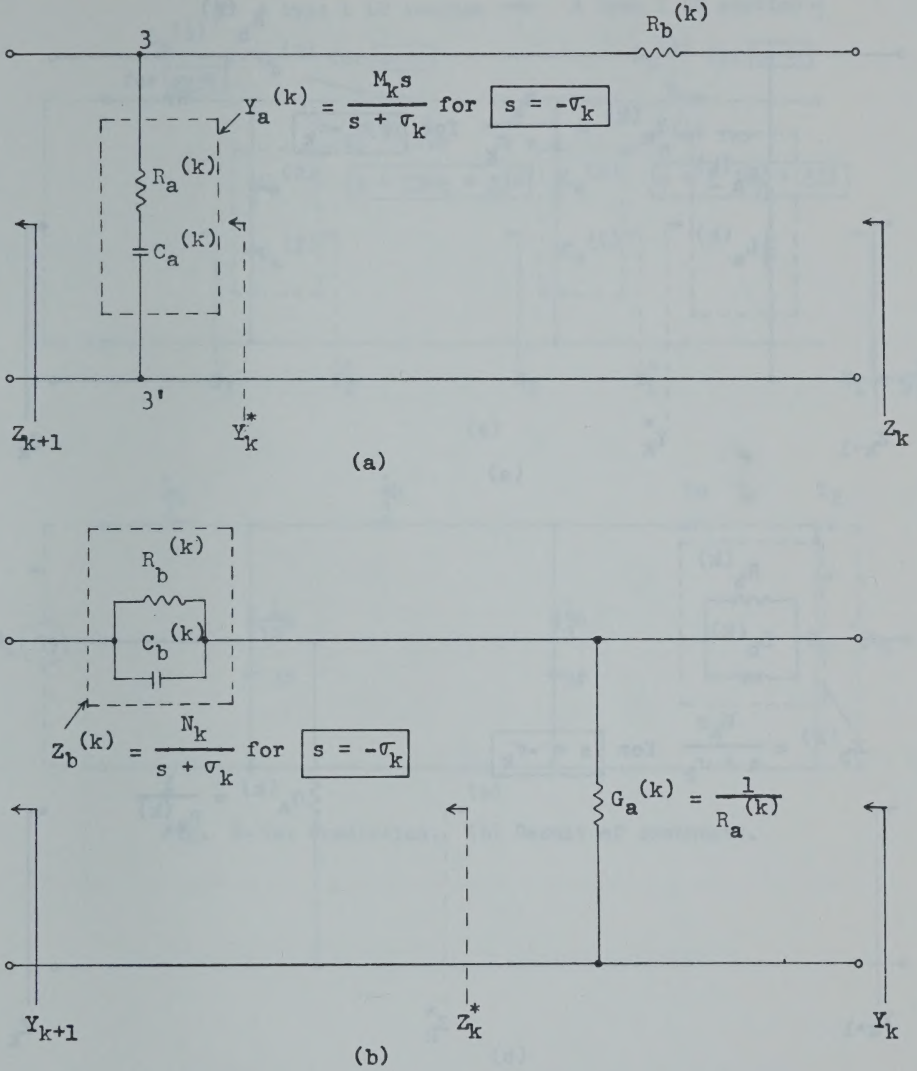

$Z_b^{(k)} = sL_b^{(k)}$ for $\boxed{s=\infty}$
3
$L_b^{(k)}$
or
$C_b^{(k)}$
$Y_a^{(k)} = \frac{M_k s}{s^2 + \omega_k^2}$
for $\boxed{s = \pm j\omega_k}$
$L_a^{(k)}$
$C_a^{(k)}$
$Z_b^{(k)} = \frac{1}{sC_b^{(k)}}$
for $\boxed{s=0}$
3'
Z_{k+1}
Y_k^*
Z_k
(a)
$L_b^{(k)}$
$C_b^{(k)}$
2
3
$Z_b^{(k)} = \frac{N_k s}{s^2 + \omega_k^2}$
for $\boxed{s = \pm j\omega_k}$
$Y_a^{(k)} = sC_a^{(k)}$
for $\boxed{s = \infty}$
or
$C_a^{(k)}$
$L_a^{(k)}$
$Y_a^{(k)} = 1/sL_a^{(k)}$
for $\boxed{s = 0}$
Y_{k+1}
Z_k^*
Y_k
(b)

Fig. 5-(a) Type 1 LC section. (b) Type 2 LC section.



$R_b^{(k)}$
3
$Y_a^{(k)} = \frac{M_k s}{s + \sigma_k}$ for $\boxed{s = -\sigma_k}$
$R_a^{(k)}$
$C_a^{(k)}$
3'
Z_{k+1}
Y_k^*
Z_k
(a)
$R_b^{(k)}$
$C_b^{(k)}$
$Z_b^{(k)} = \frac{N_k}{s + \sigma_k}$ for $\boxed{s = -\sigma_k}$
$G_a^{(k)} = \frac{1}{R_a^{(k)}}$
Y_{k+1}
Z_k^*
Y_k
(b)

Fig. 6- (a) Type 1 RC section. (b) Type 2 RC section.

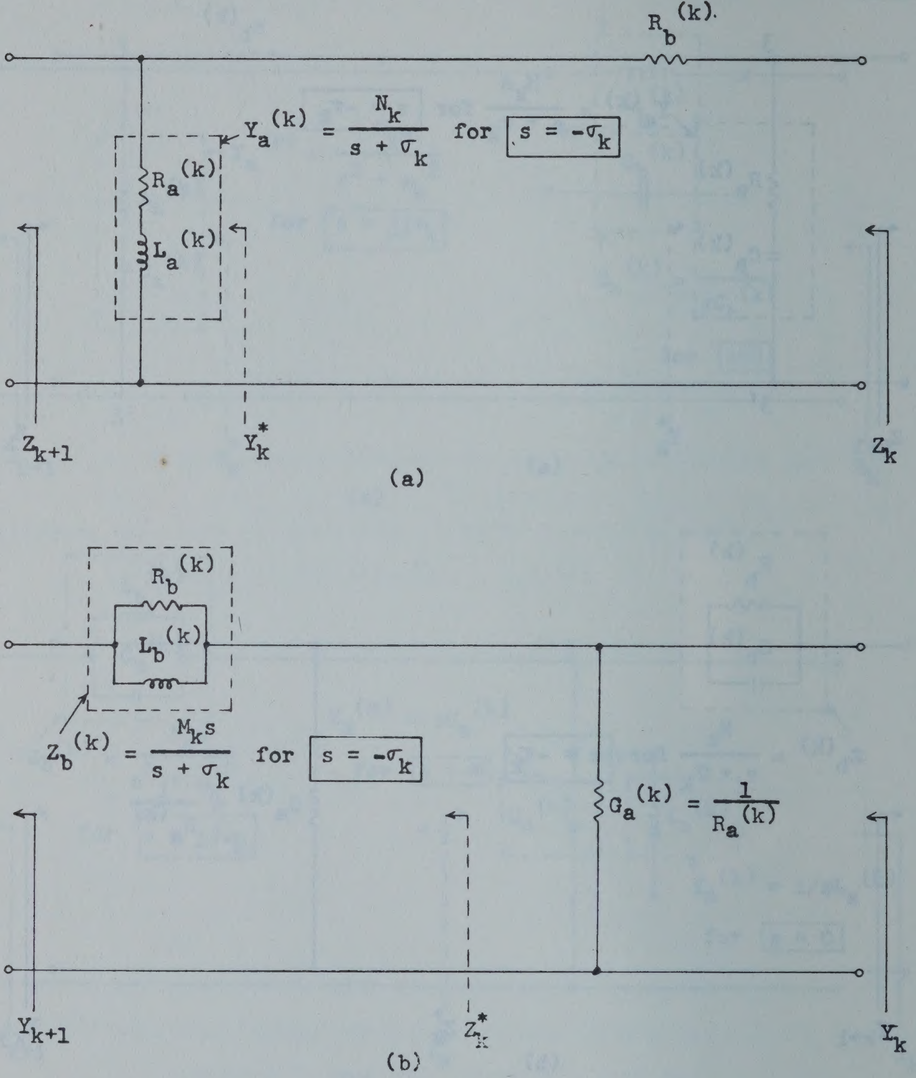

$R_b^{(k)}$
$Y_a^{(k)} = \frac{N_k}{s + \sigma_k}$ for $\boxed{s = -\sigma_k}$
$R_a^{(k)}$
$L_a^{(k)}$
Z_{k+1}
Y_k^*
Z_k
(a)
$R_b^{(k)}$
$L_b^{(k)}$
$Z_b^{(k)} = \frac{M_k s}{s + \sigma_k}$ for $\boxed{s = -\sigma_k}$
$G_a^{(k)} = \frac{1}{R_a^{(k)}}$
Y_{k+1}
Z_k^*
Y_k
(b)

Fig. 7-(a) Type 1 RL section. (b) Type 2 RL section.

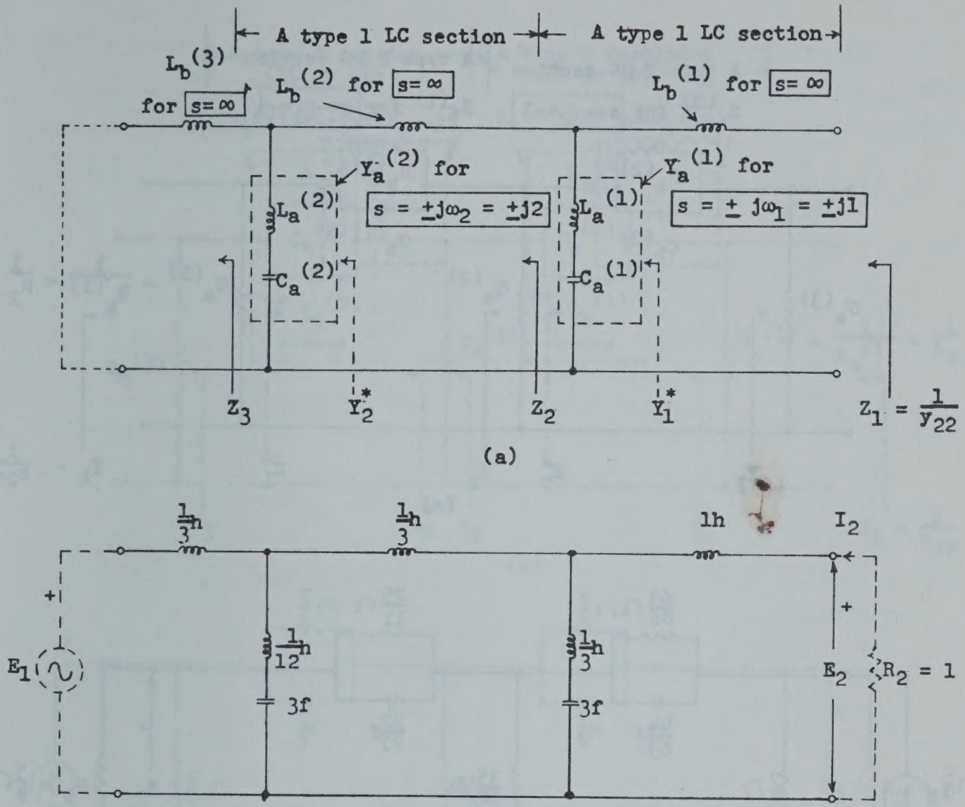

A type 1 LC section
A type 1 LC section
$L_b^{(3)}$ for $\boxed{s=\infty}$
$L_b^{(2)}$ for $\boxed{s=\infty}$
$L_b^{(1)}$ for $\boxed{s=\infty}$
$Y_a^{(2)}$ for
$\boxed{s = \pm j\omega_2 = \pm j2}$
$Y_a^{(1)}$ for
$\boxed{s = \pm j\omega_1 = \pm j1}$
$L_a^{(2)}$
$C_a^{(2)}$
$L_a^{(1)}$
$C_a^{(1)}$
Z_3
Y_2^*
Z_2
Y_1^*
$Z_1 = \frac{1}{y_{22}}$
(a)
$\frac{1}{3}$h
$\frac{1}{3}$h
1h
I_2
+
E_1
$\frac{1}{12}$h
3f
$\frac{1}{3}$h
3f
+
E_2
$R_2 = 1$
(b)

Fig. 8- (a) Prediction. (b) Result of synthesis.

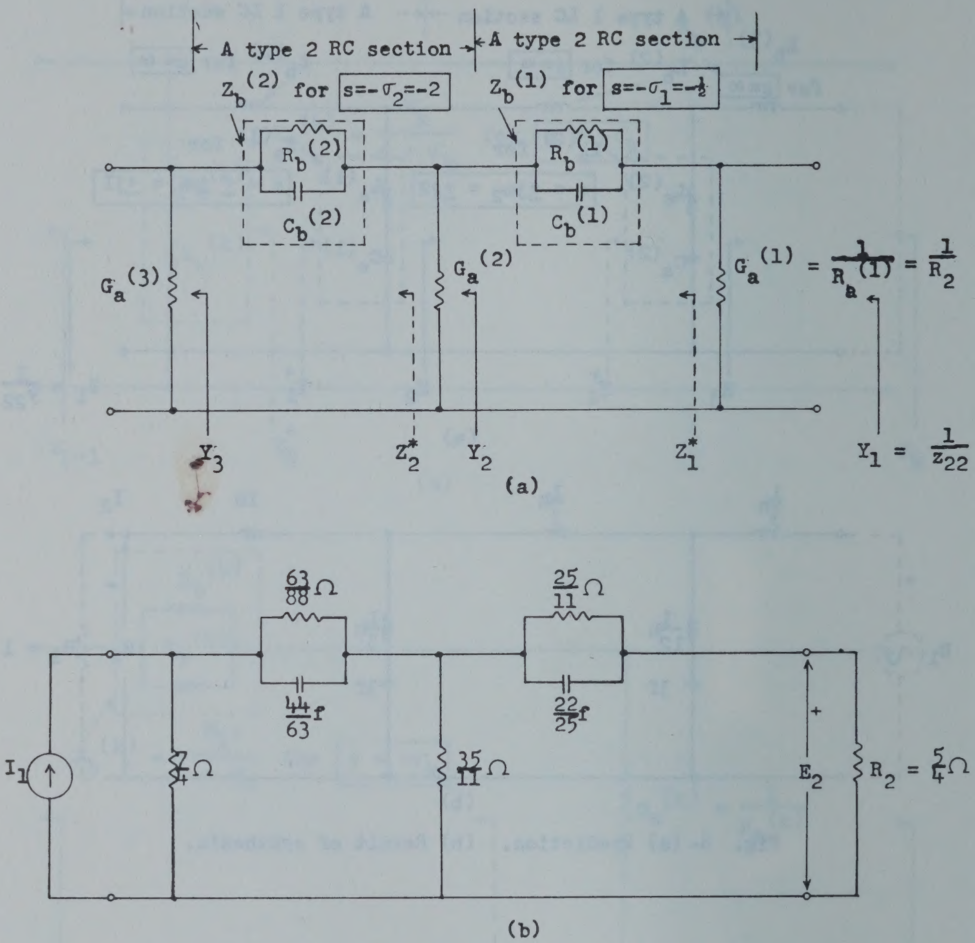

A type 2 RC section
A type 2 RC section
$Z_b^{(2)}$ for $\boxed{s=-\sigma_2=-2}$
$Z_b^{(1)}$ for $\boxed{s=-\sigma_1=-\frac{1}{2}}$
$R_b^{(2)}$
$C_b^{(2)}$
$R_b^{(1)}$
$C_b^{(1)}$
$G_a^{(3)}$
$G_a^{(2)}$
$G_a^{(1)} = \frac{1}{R_a^{(1)}} = \frac{1}{R_2}$
Y_3
Z_2^*
Y_2
Z_1^*
$Y_1 = \frac{1}{z_{22}}$
(a)
$\frac{63}{88}\Omega$
$\frac{44}{63}f$
$\frac{25}{11}\Omega$
$\frac{22}{25}f$
I_1
$\frac{7}{4}\Omega$
$\frac{35}{11}\Omega$
+
E_2
$R_2 = \frac{5}{4}\Omega$
(b)

Fig. 9-(a) Prediction. (b) Result of synthesis.

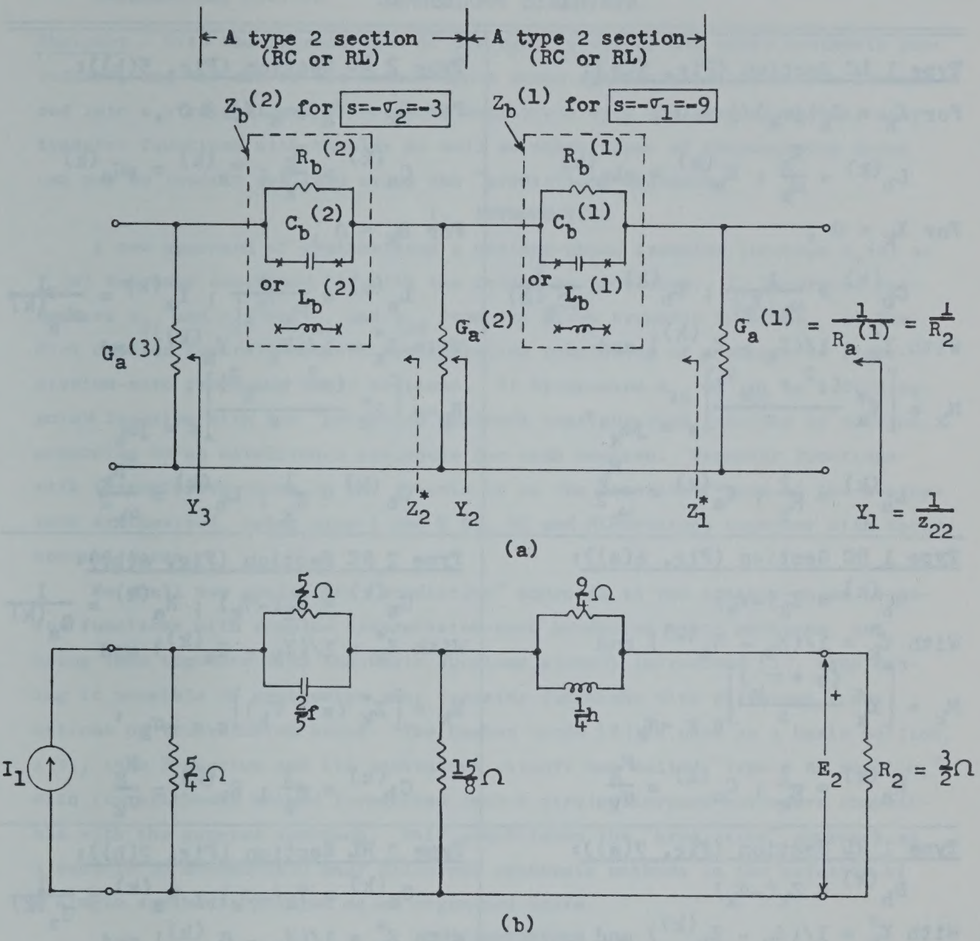

A type 2 section
(RC or RL)
A type 2 section
(RC or RL)
$Z_b^{(2)}$ for $s=-\sigma_2=-3$
$Z_b^{(1)}$ for $s=-\sigma_1=-9$
$R_b^{(2)}$
$C_b^{(2)}$
or $L_b^{(2)}$
$R_b^{(1)}$
$C_b^{(1)}$
or $L_b^{(1)}$
$G_a^{(3)}$
$G_a^{(2)}$
$G_a^{(1)} = \frac{1}{R_a^{(1)}} = \frac{1}{R_2}$
Y_3
Z_2^*
Y_2
Z_1^*
$Y_1 = \frac{1}{z_{22}}$
(a)
$\frac{5}{6}\Omega$
$\frac{2}{5}$f
$\frac{9}{4}\Omega$
$\frac{1}{4}$h
I_1
$\frac{5}{4}\Omega$
$\frac{15}{8}\Omega$
+
E_2
$R_2 = \frac{3}{2}\Omega$
(b)

Fig. 10-(a) Prediction. (b) Result of synthesis.

TABLE I

SYNTHESIS PROCEDURES

Type 1 LC Section (Fig. 5(a)): For $X_k = Z_k(j\omega_k)/j \geq 0$, $L_b^{(k)} = \frac{X_k}{\omega_k}$; $Z_b^{(k)} = sL_b^{(k)}$ For $X_k < 0$, $C_b^{(k)} = \frac{1}{\omega_k \lvert X_k \rvert}$; $Z_b^{(k)} = \frac{1}{sC_b^{(k)}}$ With $Y_k^* = 1/(Z_k - Z_b^{(k)})$ and $M_k = \left[Y_k^* \frac{(s^2 + \omega_k^2)}{s} \right]_{s = j\omega_k}$, $L_a^{(k)} = \frac{1}{M_k}$; $C_a^{(k)} = \frac{M_k}{\omega_k^2}$	Type 2 LC Section (Fig. 5(b)): For $B_k = Y_k(j\omega_k)/j \geq 0$, $C_a^{(k)} = \frac{B_k}{\omega_k}$; $Y_a^{(k)} = sC_a^{(k)}$ For $B_k < 0$, $L_a^{(k)} = \frac{1}{\omega_k \lvert B_k \rvert}$; $Y_a^{(k)} = \frac{1}{sL_a^{(k)}}$ With $Z_k^* = 1/(Y_k - Y_a^{(k)})$ and $N_k = \left[Z_k^* \frac{(s^2 + \omega_k^2)}{s} \right]_{s = j\omega_k}$, $C_b^{(k)} = \frac{1}{N_k}$; $L_b^{(k)} = \frac{N_k}{\omega_k^2}$
Type 1 RC Section (Fig. 6(a)): $R_b^{(k)} = Z_k(-\sigma_k)$ With $Y_k^* = 1/(Z_k - R_b^{(k)})$ and $M_k = \left[Y_k^* \frac{(s + \sigma_k)}{s} \right]_{s = -\sigma_k}$ $R_a^{(k)} = \frac{1}{M_k}$; $C_a^{(k)} = \frac{M_k}{\sigma_k}$	Type 2 RC Section (Fig. 6(b)): $G_a^{(k)} = Y_k(-\sigma_k)$; $R_a^{(k)} = \frac{1}{G_a^{(k)}}$ With $Z_k^* = 1/(Y_k - G_a^{(k)})$ and $N_k = \left[Z_k^* \cdot (s + \sigma_k) \right]_{s = -\sigma_k}$, $C_b^{(k)} = \frac{1}{N_k}$; $R_b^{(k)} = \frac{N_k}{\sigma_k}$
Type 1 RL Section (Fig. 7(a)): $R_b^{(k)} = Z_k(-\sigma_k)$ With $Y_k^* = 1/(Z_k - Z_b^{(k)})$ and $N_k = \left[Y_k^* \cdot (s + \sigma_k) \right]_{s = -\sigma_k}$, $L_a^{(k)} = \frac{1}{N_k}$; $R_a^{(k)} = \frac{\sigma_k}{N_k}$	Type 2 RL Section (Fig. 7(b)): $G_a^{(k)} = Y_k(-\sigma_k)$; $R_a^{(k)} = \frac{1}{G_a^{(k)}}$ With $Z_k^* = 1/(Y_k - G_a^{(k)})$ and $M_k = \left[Z_k^* \frac{(s + \sigma_k)}{s} \right]_{s = -\sigma_k}$, $R_b^{(k)} = M_k$; $L_b^{(k)} = \frac{M_k}{\sigma_k}$

SYNTHESIS OF MINIMUM-PHASE TRANSFER FUNCTIONS HAVING COMPLEX TRANSMISSION ZEROS WITH LADDER NETWORKS

WAYNE H. CHEN
University of Florida
Gainesville, Florida

Abstract.- With the introduction of new basic sections and their synthesis procedures, and the use of the Dasher cycle whose synthesis method is now formalized into a straightforward procedure compatible with the general method, many transfer functions with complex as well as other types of transmission zeros can now be readily realized using the "prediction" approach.

I. INTRODUCTION

A new approach of synthesizing a minimum-phase transfer function $Z_T(s)$ or $Y_T(s)$ has been described [1] with the following procedure: 1) Determine parameters z_{21} and z_{22} or y_{21} and y_{22} from the given transfer function. 2) Predict one (or several) network configuration consisting of a number of transmission-zero producing basic sections. 3) Synthesize z_{22} or y_{22} as a driving-point function with the "predicted" network configuration, section by section, according to an established procedure for each section. Transfer functions with transmission zeros on the $j\omega$-axis or on the negative σ-axis of the s-plane were synthesized, using type 1 and 2 LC, RC and RL sections together with their special cases.

We shall now apply this "prediction" approach to the minimum-phase transfer functions with complex transmission-zero producing basic sections, and using them together with the basic sections already introduced [1], thus making it possible to synthesize many transfer functions with different combinations of transmission zeros. The Dasher cycle [2] is used as a basic section, i.e., type 3 section and its equivalent circuit now called "type 3 RC section," with its synthesis method formalized into a straightforward procedure compatible with the general approach. This establishes the "prediction" approach as a vehicle to accommodate many different synthesis methods in the solution of a single synthesis problem on an organized basis.

II. PROTOTYPES

A. Type 1 and 2 Basic Sections

The type 1 and 2 basic sections, used earlier [1] for purely imaginary and real negative transmission zeros, are now used also to produce complex transmission zeros. These two prototypes are depicted in Fig. 1(a) and (b) with a shaded box as its "principal branch" producing one or a pair of transmission zeros (in general, at $s \neq 0$ and $s \neq \infty$). The "auxiliary branch," depicted by a clear box in the figure, contributes to a transmission zero at $s = \infty$ or $s = 0$, or does not contribute at all.

B. Type 3 and 4 Basic Sections

The type 3 and 4 basic sections are introduced in Fig. 1(c) and (d). The "principal branch," depicted with a shaded box in Fig. 1(c) or (d), produces a pair of complex transmission zeros, $s = -\alpha \pm j\beta$ as expected, because Z_B as well as Y_B has them as poles [1]; the "auxiliar branches" depicted with clear boxes produce no transmission zeros (even though Y_A and Y_C as well as Z_A and Z_C have $s = -\gamma$ as poles) under the condition:

$$B = \frac{A + C}{AC} \; \frac{\gamma^2 - 2\alpha\gamma^2 + \alpha^2 + \beta^2}{\gamma}. \tag{1}$$

We shall assume that the above condition is always imposed in the discussion throughout this paper.

The above properties may be readily established by using a type 3 (or 4) basic section in a ladder network, obtaining its z_{21} (or y_{21}) parameter and examining its poles which are transmission zeros.

III. BASIC SECTIONS

A number of basic sections belonging to the four prototypes described above are introduced in the following paragraphs.

A. Type 1 RLC Section

Fig. 2 depicts a type 1 RLC section whose principal branch $Y_a^{(k)}$ is responsible for a pair of complex conjugate transmission zeros $s = -\alpha_k \pm j\beta_k$, and whose auxiliary branch $Z_b^{(k)}$ contributes to a transmission zero at $s = \infty$ or $s = 0$. The super- or subscript k indicates that this type 1 RLC section is used as the kth transmission-zero-producing section of a network.

The procedure of synthesizing a given Z_k with a type 1 RLC section is given in Table I. This procedure may be readily justified in this manner: At the transmission zero $s = -\alpha_k + j\beta_k$, we have $Z_a^{(k)} = 1/Y_a^{(k)} = 0$, i.e., a short circuit across 3 and 3' in Fig. 4, and $[Z_k]_{s = -\alpha_k + j\beta_k} = [Z_b^{(k)}]_{-\alpha_k + j\beta_k}$. The latter implies that, for $Z_k(-\alpha_k + j\beta_k) = R_k + jX_k$,

$$R_k + jX_k = (-\alpha_k + j\beta_k)L_b^{(k)} + R_b^{(k)}, \quad X_k \geq 0, \tag{2}$$

$$R_k + jX_k = \frac{1}{(-\alpha_k + j\beta_k)C_b^{(k)}} + R_b^{(k)}, \quad X_k < 0, \tag{3}$$

from which we obtain the expressions for $L_b^{(k)}$, $R_b^{(k)}$ and $C_b^{(k)}$, $R_b^{(k)}$ in Table I. With $Z_b^{(k)}$ in Fig. 2 now evaluated, we may partial-fraction $Y_k^* = 1/(Z_k - Z_b^{(k)})$ in the form of

$$Y_k^* = Y_a^{(k)} + \frac{1}{Z_{k+1}} = \frac{M_k(s + \gamma_k)}{s^2 + 2\alpha_k s + \alpha_k^2 + \beta_k^2} + \frac{1}{Z_{k+1}}. \tag{4}$$

By identifying $Y_a^{(k)} = M_k(s + \gamma_k)/(s^2 + 2\alpha_k s + \alpha_k^2 + \beta_k^2)$ with the admittance of the $Y_a^{(k)}$ branch in Fig. 2, we obtain the expressions for $L_a^{(k)}$, $R_a^{(k)}$, $C_a^{(k)}$ and $R_C^{(k)}$ in Table I.

B. Type 2 RLC Section

Fig. 3 depicts a type 2 RLC section whose principal branch $Z_b^{(k)}$ is responsible for a pair of complex conjugate transmission zeros $s = -\alpha_k \pm j\beta_k$, and whose auxiliary branch $Y_a^{(k)}$ contributes to a transmission zero at $s = \infty$ or $s = 0$. The procedure of synthesizing a given Z_k with a type 2 RLC section is tabulated in Table I, and may be readily justified in the same manner as with a type 1 RLC section (Section III-A).

C. Type 3 Sections

Fig. 4(a) depicts a type 3 section together with a parallel resistance $R_a^{(h)}$ which often comes along as the result of synthesis. This type 3 section has the circuits in Fig. 5, i.e., a "type 3 RC section" in Fig. 5(a) or (b) and a "type 3 RLC section" in Fig. 5(c), as its equivalent circuits.

We shall now give the synthesis procedure for a type 3 section together with a parallel resistance $R_a^{(h)}$. Letting $s = -\alpha_k \pm j\beta_k = -\alpha_h \pm j\beta_h$, we define:

$$[Y_h]_{s = -\alpha_h + j\beta_h} = g_h + jb_h, \quad \left[\frac{d}{ds} Y_h\right]_{s = -\alpha_h + j\beta_h} = g'_h + jb'_h \tag{5}$$

$$g_h^* = b_h \frac{b_h - \beta_h g'_h + \alpha_h b'_h}{(\alpha_h/\beta_h) b_h - \alpha_h g'_h - \beta_h b'_h}. \tag{6}$$

If $g_h < g_h^*$, we cannot synthesize Y_h with a type 3 section as shown in Fig. 4(a). For $g_h \geq g_h^*$, we proceed as follows:

Pull out in Fig. 4(a),

$$G_a^{(h)} = \frac{1}{R_a^{(h)}} = g_h - g_h^*. \tag{7}$$

For $Y_k = Y_h - G_a^{(h)}$ and $Y_k(-\alpha_k + j\beta_k) = g_k + jb_k$, we compute

$$\gamma_k = \frac{1}{\beta_k} \frac{\alpha_k^2 + \beta_k^2}{(g_k/b_k) + (\alpha_k/\beta_k)}, \quad A_k = \frac{1}{b_k} \frac{g_k^2 + b_k^2}{(g_k/b_k) + (\alpha_k/\beta_k)} \tag{8}$$

and $Y_A^{(k)}$ in Fig. 4(a) is now a known expression. With $Z_k^* = 1/(Y_k - Y_A^{(k)})$, we evaluate

$$B_k = \left[Z_k^* \frac{s^2 + 2\alpha_k s + \alpha_k^2 + \beta_k^2}{s + \gamma_k} \right]_{s = -\alpha_k + j\beta_k} \tag{9}$$

and $Z_B^{(k)}$ in Fig. 4(a) is now a known quantity. With $Y_k^{**} = 1/(Z_k^* - Z_B^{(k)})$, we compute

$$C_k = \left[Y_k^{**} \frac{s + \gamma_k}{s} \right]_{s = -\alpha_k + j\beta_k} \tag{10}$$

and $Y_C^{(k)}$ is now a known expression. With $Y_A^{(k)}$, $Z_B^{(k)}$ and $Y_C^{(k)}$ all known in Fig. 4(a), we replace this "evaluated" type 3 section with one of its equivalent circuits in Fig. 5. We are now ready to synthesize $Y_{k+1} = Y_k^{**} - Y_C^{(k)}$ in Fig. 4(a) for the next section.

The above procedure may be justified in this manner: By pulling out a conductance $G_a^{(h)} = g_h - g_h^*$, where g_h^* is defined in (6), Dasher [2] established that the constant B_k in the $Z_B^{(k)}$ expression in Fig. 4(a) is real, condition (1) is satisfied, and $Y_k = Y_h - G_a^{(h)}$ is realizable with a type 3 section. To evaluate $Y_A^{(k)}$, $Z_B^{(k)}$ and $Y_C^{(k)}$ is to obtain appropriate partial fractions from Y_k, Z_k^* and Y_k^{**}.

D. Type 4 Sections

Fig. 4(b) depicts a type 4 section together with a series resistance $R_b^{(h)}$ which often comes along as the result of synthesis. This type 4 section has the circuits in Fig. 6, i.e., a "type 4 RL section" in Fig. 6(a) or (b) and a "type 4 RLC section" in Fig. 6(c), as its equivalent circuits.

We shall now give the synthesis procedure for a type 4 section together with a series resistance $R_b^{(h)}$. Letting $s = -\alpha_k \pm j\beta_k = -\alpha_h \pm j\beta_h$, we define:

$$\Big[Z_h\Big]_{s = -\alpha_h + j\beta_h} = r_h + jx_h, \quad \Big[\frac{d}{ds} Z_h\Big]_{s = -\alpha_h + j\beta_h} = r_h' + jx_h', \tag{11}$$

$$r_h^* = x_h \frac{x_h - \beta_h r_h' + \alpha_h x_h'}{(\alpha_h/\beta_h)x_h - \alpha_h r_h' - \beta_h x_h'} . \tag{12}$$

If $r_h < r_h^*$, we cannot synthesize Z_h with a type 4 section as shown in Fig. 4(b). For $r_h \geq r_h^*$, we proceed as follows:

Pull out in Fig. 4(b)

$$R_b^{(h)} = r_h - r_h^* \tag{13}$$

For $Z_k = Z_h - R_b^{(h)}$ and $Z_k(-\alpha_k + j\beta_k) = r_k + jx_k$, we compute

$$\gamma_k = \frac{1}{\beta_k} \frac{\alpha_k^2 + \beta_k^2}{(r_k/x_k) + (\alpha_k/\beta_k)} , \quad A_k = \frac{1}{x_k} \frac{r_k^2 + x_k^2}{(r_k/x_k) + (\alpha_k/\beta_k)} \tag{14}$$

and $Z_A^{(k)}$ in Fig. 4(b) is now a known expression. With $Y_k^* = 1/(Z_k - Z_A^{(k)})$, we evaluate

$$B_k = \left[Y_k^* \frac{s^2 + 2\alpha_k s + \alpha_k^2 + \beta_k^2}{s + \gamma_k} \right]_{s = -\alpha_k + j\beta_k} \tag{15}$$

and $Y_B^{(k)}$ in Fig. 4(b) is now a known expression. With $Z_k^{**} = 1/(Y_k^* - Y_B^{(k)})$, we compute

$$C_k = \left[Z_k^{**} \frac{s + \gamma_k}{s} \right]_{s = -\alpha_k + j\beta_k} \qquad (16)$$

and $Z_C^{(k)}$ is now a known expression. With $Z_A^{(k)}$, $Y_B^{(k)}$ and $Z_C^{(k)}$ all known in Fig. 4(b), we replace the "evaluated" type 4 section with one of its equivalent circuits in Fig. 6. We are now ready to synthesize $Z_{k+1} = Z_k^{**} - Z_C^{(k)}$ in Fig. 4(b) for the next section.

The above procedure may be justified in the same manner as that for a type 3 section (Section III-C).

IV. PREDICTION

For reasons already established [1] we predict the network, for a given distribution of transmission zeros, in the following manner: 1) Fit it into a standard ladder form with two parallel (or series) branches at both ends for prescribed z_{21} and z_{22} (or y_{21} and y_{22}). 2) Use the principal branches of the basic sections to produce the finite non-zero-value transmission zeros, permitting some modification for special cases. 3) Use the auxiliary and/or additional branches to contribute to the transmission zeros at $s = 0$ or $s = \infty$, or to make no contribution if the distribution has no transmission zero at $s = 0$ and $s = \infty$.

Again for reasons already established [1] and modified here, the following rules are often found to be useful: For prescribed z_{21} and z_{22} (or y_{21} and y_{22}), it is often desirable to use a network configuration with parallel (or series) resistances at both ends to simulate the source and load impedance (or admittance). For simple orderly development, such a configuration may consist exclusively of type 2 (or 1) sections together with some type 3 (or 4) sections and a parallel (or series) branch, which is either a resistance itself or has a parallel (or series) resistive component, on the extreme left.

V. SYNTHESIS PROBLEMS

A. Illustration of An RC or RLC Synthesis Problem

$$\text{For } Z_T(s) = H\,\frac{P(s)}{Q(s)} = H\,\frac{(s + 1/2)(s^2 + 5s + 13/2)}{(s + 1)(s + 2)(s + 3)}\,, \qquad (17)$$

we obtain [1]

$$z_{21} = H\,\frac{P(s)}{Q(s)}\,,\quad z_{22} = \frac{(2s + 3)(2s + 5)(2s + 7)}{Q(s)} \qquad (18)$$

with z_{22} chosen for RC synthesis. With transmission zeros $s = (-5\pm j)/2$ and $s = -1/2$, we predict the configuration in Fig. 7(a). Using the procedures for type 3 section (Section III-C) and type 2 RC section [1], we obtain the preliminary result in Fig. 7(b). Replacing the type 3 section with its equivalent type 3 RC section in Fig. 5(a) and pulling R_2 out as the load [1], the RC ladder arrangement in Fig. 7(c) is obtained. We may also obtain an RLC arrangement if Fig. 5(c) is instead used.

B. Illustration of An RL or RLC Synthesis Problem

For

$$Y_T(s) = -H\,\frac{P(s)}{Q(s)} = -H\,\frac{(s + 1/2)(s^2 + 5s + 13/2)}{(s + 1)(s + 2)(s + 3)}\,, \qquad (19)$$

we obtain [1]

$$y_{21} = -H\,\frac{P(s)}{Q(s)}\,,\; y_{22} = \frac{(2s + 3)(2s + 5)(2s + 7)}{Q(s)} \qquad (20)$$

with y_{22} chosen for RL synthesis. With transmission zeros, $s = -(5 \pm j)/2$ and $s = -1/2$, we predict the configuration in Fig. 8(a), and synthesize to obtain Fig. 8(b) and then Fig. 8(c) with the aid of Fig. 6(a). We may also obtain an RLC arrangement, if Fig. 6(c) is instead used.

C. Illustration of An RLC Synthesis Problem

For

$$Y_T(s) = -H\,\frac{P(s)}{Q(s)} = -H\,\frac{(s + 4)(s^2 + 4s + 8)}{s^4 + 13s^3 + 60s^2 + 126s + 104}\,, \qquad (21)$$

we obtain [1]

$$y_{21} = -H\,\frac{P(s)}{Q(s)}\,,\; y_{22} = \frac{s^3 + 8s^2 + 28s + 32}{Q(s)} \qquad (22)$$

with y_{22} chosen to be a realizable driving-point function. With transmission zeros $s = -2 \pm j2$, $s = -4$ and $s = \infty$, we predict the configuration in Fig. 9(a) with a type 1 RLC section and a type 1 section, the latter being an RC or RL section to be determined for realizability by the synthesis procedures. Since $R_b^{(2)}$ alone of a type 1 section does not contribute to the transmission zero $s = \infty$ and we may use "auxiliary and/or additional branches" (Section IV) to contribute to this $s = \infty$, $L_b^{(2)}$ is the additional branch for this purpose. Note that $Z_b^{(1)}$, $sL_b^{(2)} + R_b^{(2)}$ and $Z_b^{(3)}$ all contribute to $s = \infty$, and $Z_b^{(3)}$ being a series branch on the extreme left has a series resistive component $R_b^{(3)}$. Fig. 9(b) depicts the result of synthesis.

The straightforwardness of the synthesis methods as demonstrated in the preceding illustrations also suggests the possibility that the "prediction" approach may provide a basis for computer design. For example, the computer may be programmed to try out 1) different orders of the transmission zeros to be realized with a predicted configuration, or 2) even different possible predictions, in the synthesis of a given transfer function.

BIBLIOGRAPHY

[1] W. H. Chen, "A New Approach in Transfer-function Synthesis," Proceedings of the 1963 Allerton Conference on Circuit and System Theory (November 15-17, University of Illinois).

[2] B. J. Dasher, "Synthesis of RC transfer function as unbalanced two terminal-pair networks," IRE Trans. on Circuit Theory, vol. CT-1, pp. 20-34; December, 1952.

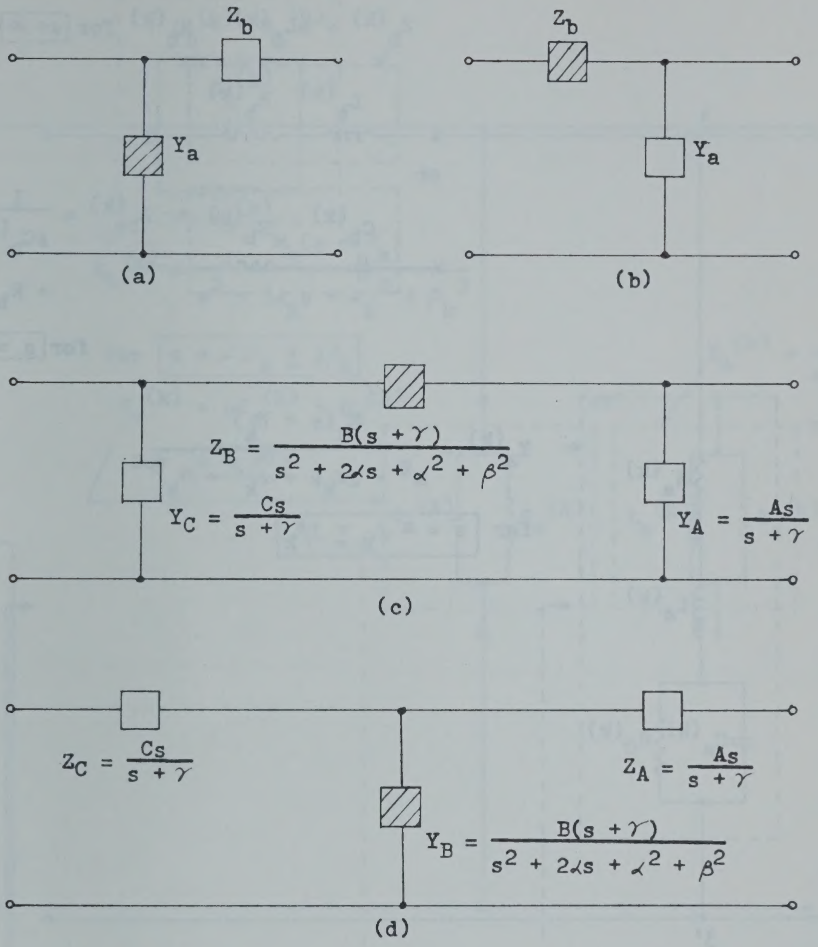

Z_b
Y_a
(a)
Z_b
Y_a
(b)
$Z_B = \frac{B(s + \gamma)}{s^2 + 2\alpha s + \alpha^2 + \beta^2}$
$Y_C = \frac{Cs}{s + \gamma}$
$Y_A = \frac{As}{s + \gamma}$
(c)
$Z_C = \frac{Cs}{s + \gamma}$
$Z_A = \frac{As}{s + \gamma}$
$Y_B = \frac{B(s + \gamma)}{s^2 + 2\alpha s + \alpha^2 + \beta^2}$
(d)

Fig. 1 Prototype basic sections with shaded boxes depicting principal branches and clear boxes, auxiliary branches. (a) Type 1 section. (b) Type 2 section. (c) Type 3 section. (d) Type 4 section.

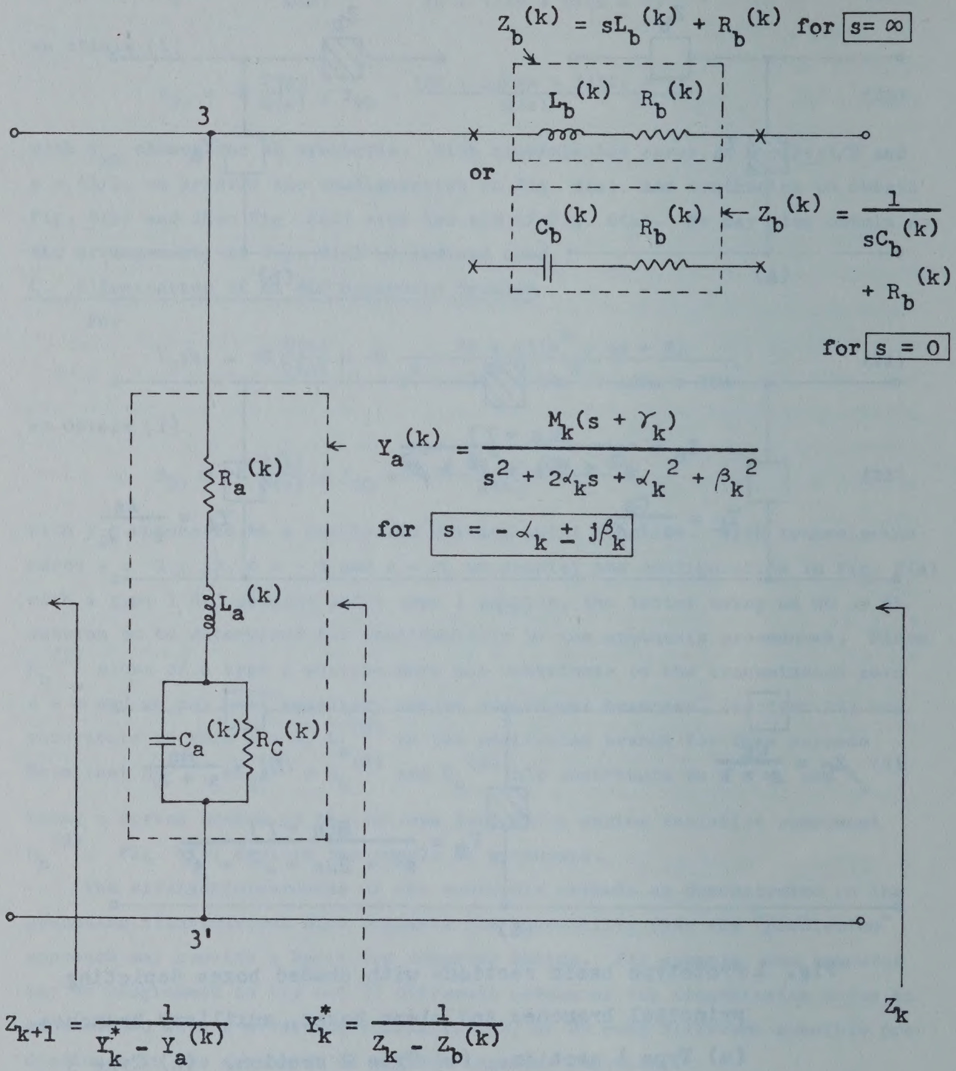

$Z_b^{(k)} = sL_b^{(k)} + R_b^{(k)}$ for $\boxed{s=\infty}$
$L_b^{(k)}$
$R_b^{(k)}$
3
or
$C_b^{(k)}$
$R_b^{(k)}$
$Z_b^{(k)} = \frac{1}{sC_b^{(k)}} + R_b^{(k)}$
for $\boxed{s = 0}$
$R_a^{(k)}$
$Y_a^{(k)} = \frac{M_k(s + \gamma_k)}{s^2 + 2\alpha_k s + \alpha_k^2 + \beta_k^2}$
for $\boxed{s = -\alpha_k \pm j\beta_k}$
$L_a^{(k)}$
$C_a^{(k)}$
$R_C^{(k)}$
3'
$Z_{k+1} = \frac{1}{Y_k^* - Y_a^{(k)}}$
$Y_k^* = \frac{1}{Z_k - Z_b^{(k)}}$
Z_k

Fig. 2-Type 1 RLC section.

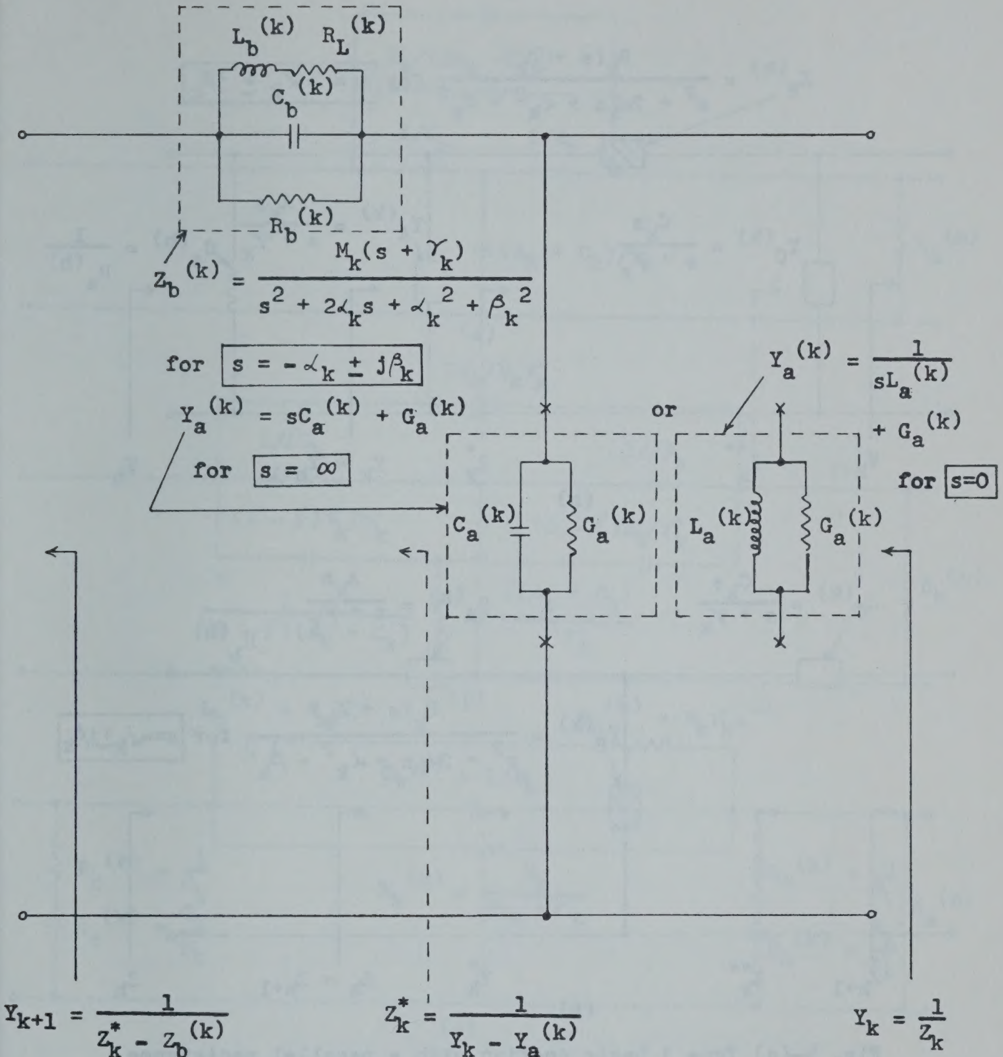

$L_b^{(k)}$
$R_L^{(k)}$
$C_b^{(k)}$
$R_b^{(k)}$
$Z_b^{(k)} = \frac{M_k(s+\gamma_k)}{s^2+2\alpha_k s+\alpha_k^2+\beta_k^2}$
for $\boxed{s = -\alpha_k \pm j\beta_k}$
$Y_a^{(k)} = sC_a^{(k)} + G_a^{(k)}$
for $\boxed{s = \infty}$
or
$Y_a^{(k)} = \frac{1}{sL_a^{(k)}} + G_a^{(k)}$
for $\boxed{s=0}$
$C_a^{(k)}$
$G_a^{(k)}$
$L_a^{(k)}$
$G_a^{(k)}$
$Y_{k+1} = \frac{1}{Z_k^* - Z_b^{(k)}}$
$Z_k^* = \frac{1}{Y_k - Y_a^{(k)}}$
$Y_k = \frac{1}{Z_k}$

Fig. 3—Type 2 RLC section.

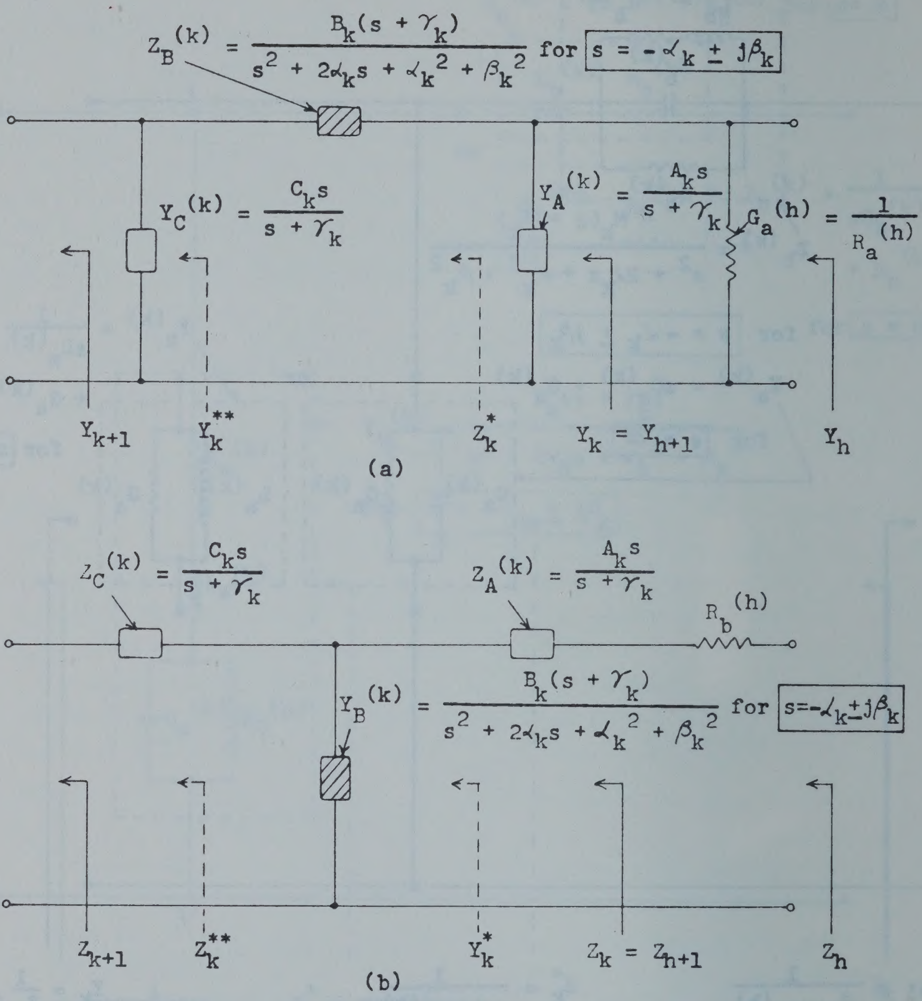

$Z_B^{(k)} = \frac{B_k(s+\gamma_k)}{s^2+2\alpha_k s+\alpha_k^2+\beta_k^2}$ for $\boxed{s=-\alpha_k \pm j\beta_k}$
$Y_C^{(k)} = \frac{C_k s}{s+\gamma_k}$
$Y_A^{(k)} = \frac{A_k s}{s+\gamma_k}$
$G_a^{(h)} = \frac{1}{R_a^{(h)}}$
Y_{k+1}
Y_k^{**}
Z_k^*
$Y_k = Y_{h+1}$
Y_h
(a)
$Z_C^{(k)} = \frac{C_k s}{s+\gamma_k}$
$Z_A^{(k)} = \frac{A_k s}{s+\gamma_k}$
$R_b^{(h)}$
$Y_B^{(k)} = \frac{B_k(s+\gamma_k)}{s^2+2\alpha_k s+\alpha_k^2+\beta_k^2}$ for $\boxed{s=-\alpha_k \pm j\beta_k}$
Z_{k+1}
Z_k^{**}
Y_k^*
$Z_k = Z_{h+1}$
Z_h
(b)

Fig. 4—(a) Type 3 basic section with a parallel resistance $R_a^{(h)}$. (b) Type 4 basic section with a series resistance $R_b^{(h)}$.

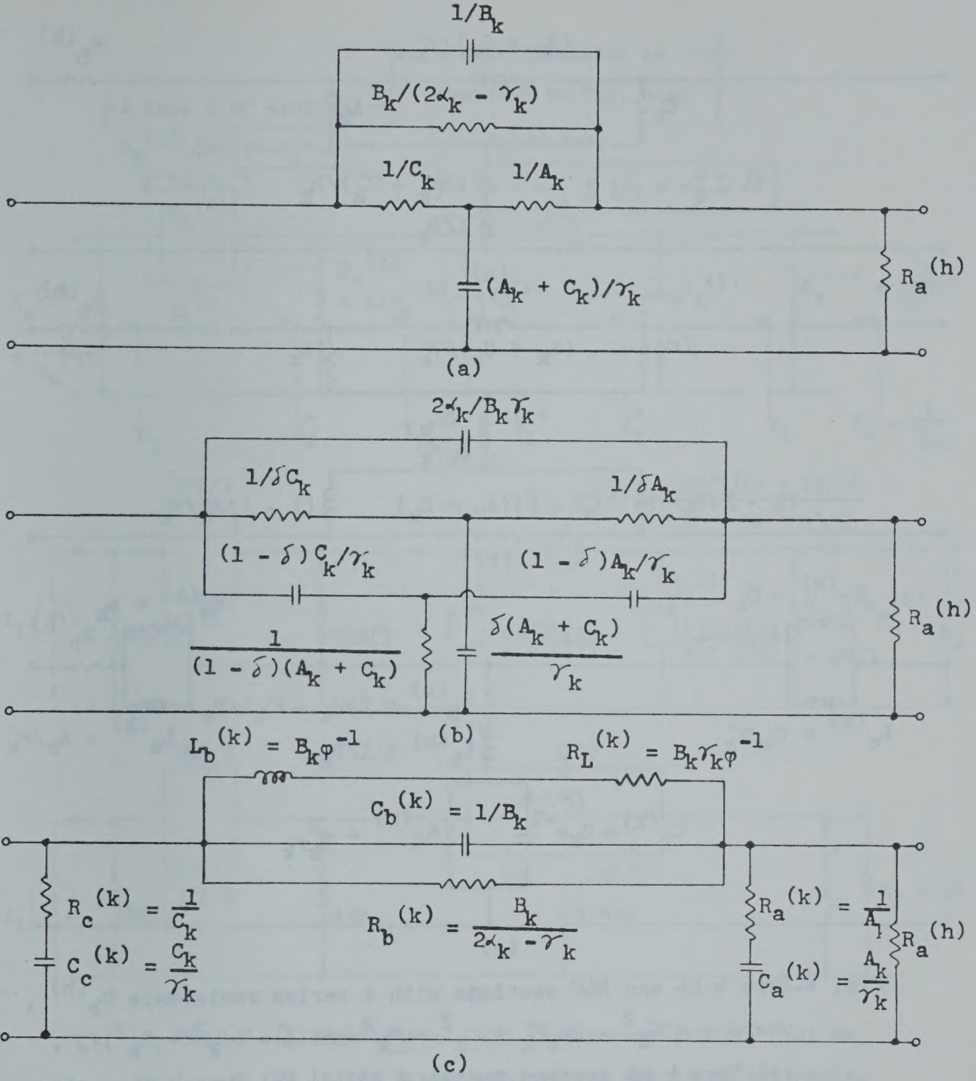

$1/B_k$
$B_k/(2\alpha_k - \gamma_k)$
$1/C_k$
$1/A_k$
$(A_k + C_k)/\gamma_k$
$R_a^{(h)}$
(a)
$2\alpha_k/B_k\gamma_k$
$1/\delta C_k$
$1/\delta A_k$
$(1 - \delta)C_k/\gamma_k$
$(1 - \delta)A_k/\gamma_k$
$\frac{1}{(1-\delta)(A_k + C_k)}$
$\frac{\delta(A_k + C_k)}{\gamma_k}$
$R_a^{(h)}$
(b)
$L_b^{(k)} = B_k\varphi^{-1}$
$R_L^{(k)} = B_k\gamma_k\varphi^{-1}$
$C_b^{(k)} = 1/B_k$
$R_c^{(k)} = \frac{1}{C_k}$
$C_c^{(k)} = \frac{C_k}{\gamma_k}$
$R_b^{(k)} = \frac{B_k}{2\alpha_k - \gamma_k}$
$R_a^{(k)} = \frac{1}{A_k}$
$C_a^{(k)} = \frac{A_k}{\gamma_k}$
$R_a^{(h)}$
(c)

Fig. 5—Type 3 RC and RLC sections with a parallel resistance $R_a^{(h)}$, where $\varphi = \gamma_k^2 - 2\alpha_k\gamma_k + \alpha_k^2 + \beta_k^2$ and $\delta = (\alpha_k^2 + \beta_k^2)/\varphi$. (a) Type 3 RC section for $2\alpha_k \geq \gamma_k$. (b) Type 3 RC section for $2\alpha_k < \gamma_k$. (c) Type 3 RLC section for $2\alpha_k \geq \gamma_k$.

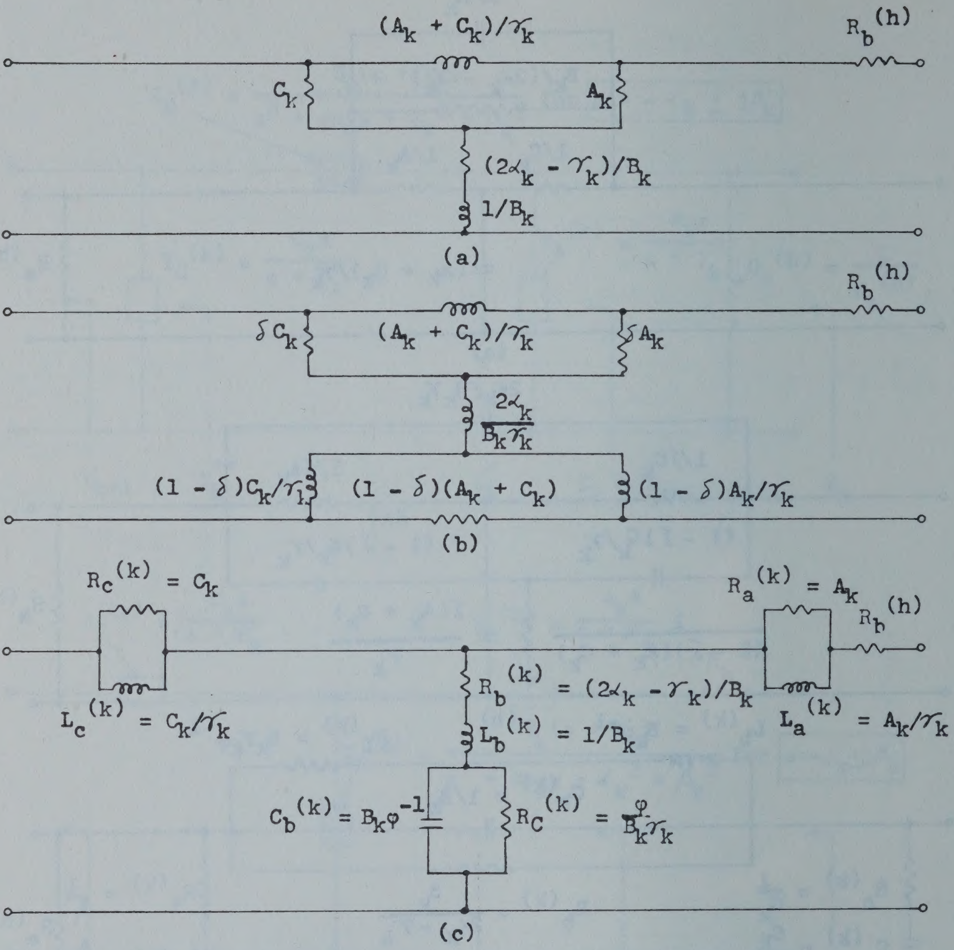

$(A_k + C_k)/\gamma_k$
$R_b^{(h)}$
C_k
A_k
$(2\alpha_k - \gamma_k)/B_k$
$1/B_k$
(a)
$R_b^{(h)}$
δC_k
$(A_k + C_k)/\gamma_k$
δA_k
$\frac{2\alpha_k}{B_k\gamma_k}$
$(1 - \delta)C_k/\gamma_k$
$(1 - \delta)(A_k + C_k)$
$(1 - \delta)A_k/\gamma_k$
(b)
$R_c^{(k)} = C_k$
$L_c^{(k)} = C_k/\gamma_k$
$R_a^{(k)} = A_k$
$R_b^{(h)}$
$L_a^{(k)} = A_k/\gamma_k$
$R_b^{(k)} = (2\alpha_k - \gamma_k)/B_k$
$L_b^{(k)} = 1/B_k$
$C_b^{(k)} = B_k\varphi^{-1}$
$R_C^{(k)} = \frac{\varphi}{B_k\gamma_k}$
(c)

Fig. 6—Type 4 RL and RLC sections with a series resistance $R_b^{(h)}$, where $\varphi = \gamma_k^2 - 2\alpha_k\gamma_k + \alpha_k^2 + \beta_k^2$ and $\delta = (\alpha_k^2 + \beta_k^2)/\delta$. (a) Type 4 RL section for $2\alpha_k \geqslant \gamma_k$. (b) Type 4 RL section for $2\alpha_k < \gamma_k$. (c) Type 4 RLC section for $2\alpha_k \geqslant \gamma_k$.

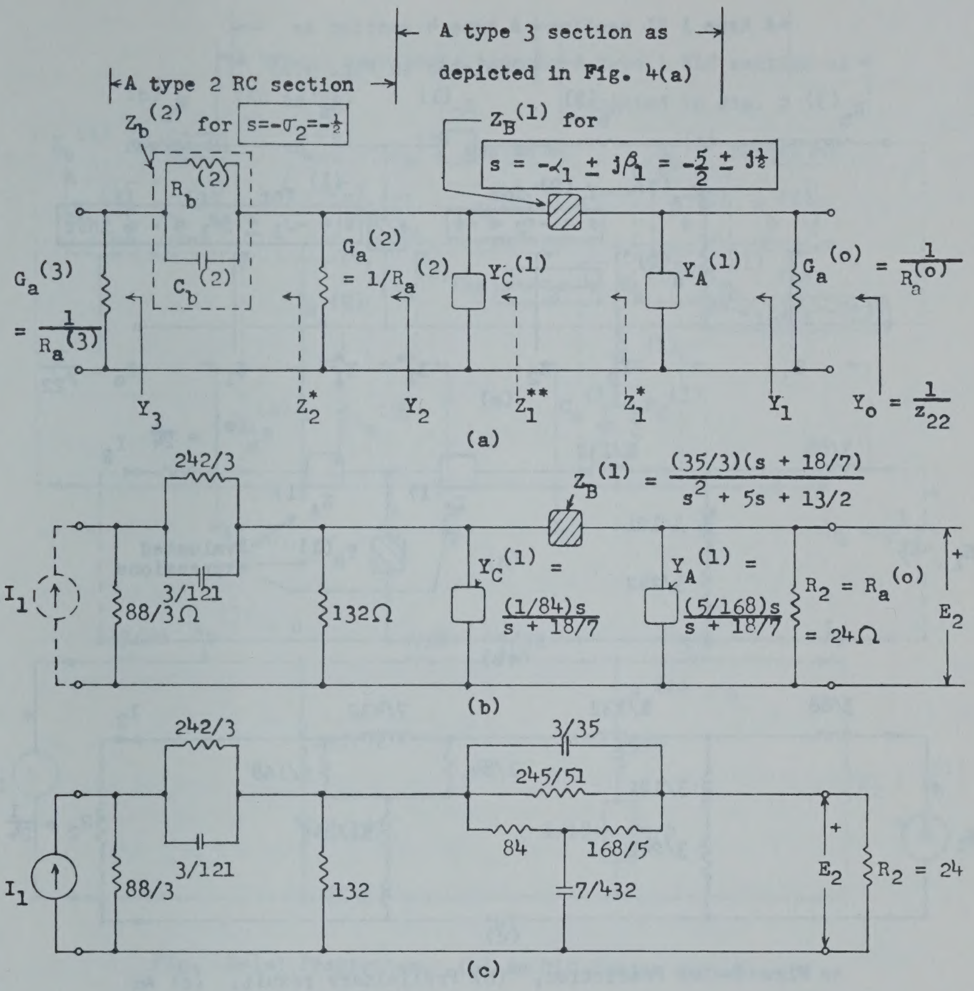

A type 3 section as depicted in Fig. 4(a)
A type 2 RC section
$Z_b^{(2)}$ for $s=-\sigma_2=-\frac{1}{2}$
$Z_B^{(1)}$ for
$s = -\alpha_1 \pm j\beta_1 = -\frac{5}{2} \pm j\frac{1}{2}$
$R_b^{(2)}$
$C_b^{(2)}$
$G_a^{(3)} = \frac{1}{R_a^{(3)}}$
$G_a^{(2)} = 1/R_a^{(2)}$
$Y_C^{(1)}$
$Y_A^{(1)}$
$G_a^{(o)} = \frac{1}{R_a^{(o)}}$
Y_3
Z_2^*
Y_2
Z_1^{**}
Z_1^*
Y_1
$Y_o = \frac{1}{z_{22}}$
(a)
242/3
3/121
88/3Ω
132Ω
$Z_B^{(1)} = \frac{(35/3)(s + 18/7)}{s^2 + 5s + 13/2}$
$Y_C^{(1)} = \frac{(1/84)s}{s + 18/7}$
$Y_A^{(1)} = \frac{(5/168)s}{s + 18/7}$
$R_2 = R_a^{(o)} = 24\Omega$
I_1
E_2
+
(b)
242/3
3/121
88/3
132
3/35
245/51
84
168/5
7/432
I_1
E_2
+
$R_2 = 24$
(c)

Fig. 7- (a) Prediction. (b) Preliminary result. (c) An RC ladder arrangement as the result of synthesis.

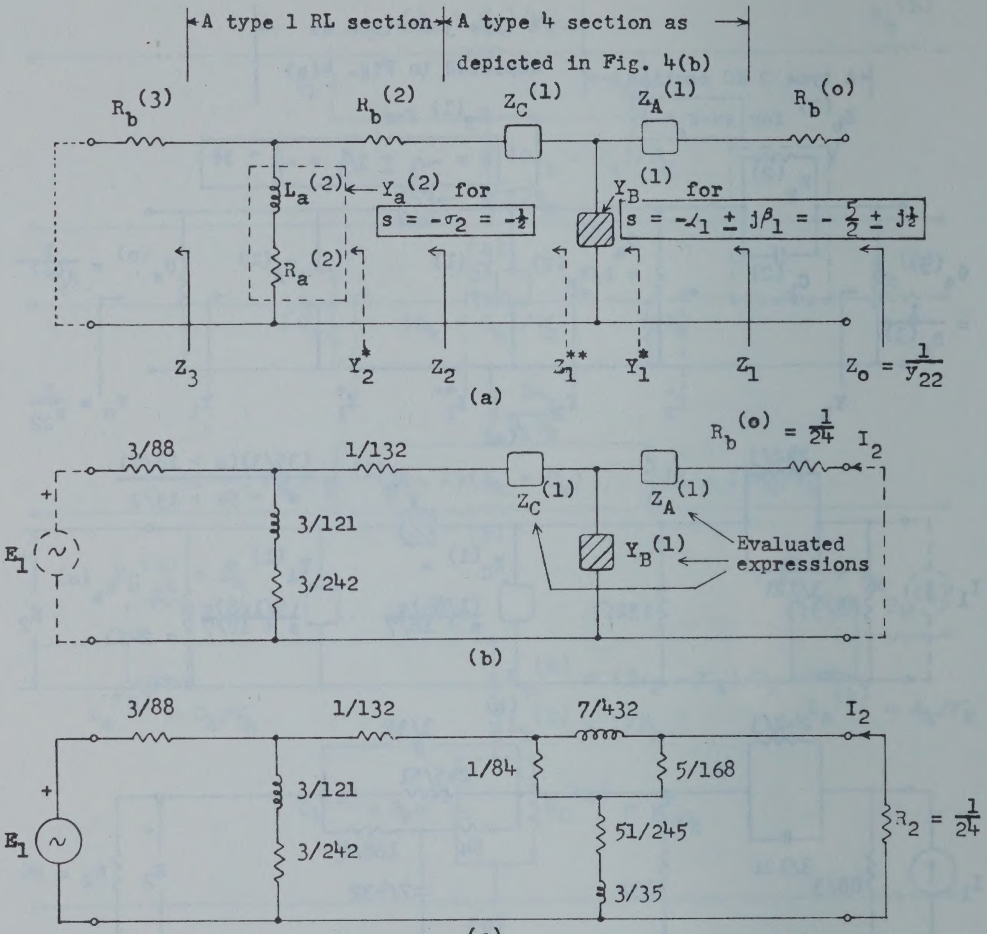

A type 1 RL section
A type 4 section as depicted in Fig. 4(b)
R_b(3)
R_b(2)
Z_C(1)
Z_A(1)
R_b(o)
L_a(2)
R_a(2)
Y_a(2) for
s = -σ2 = -½
Y_B(1) for
s = -α1 ± jβ1 = -5/2 ± j½
Z3
Y2*
Z2
Z1**
Y1*
Z1
Zo = 1/y22
(a)
3/88
1/132
R_b(o) = 1/24
I2
+
E1
3/121
3/242
Z_C(1)
Z_A(1)
Y_B(1)
Evaluated expressions
(b)
3/88
1/132
7/432
I2
1/84
5/168
+
E1
3/121
3/242
51/245
3/35
R2 = 1/24
(c)

Fig. 8-(a) Prediction. (b) Preliminary result. (c) An RL ladder arrangement as the result of synthesis.

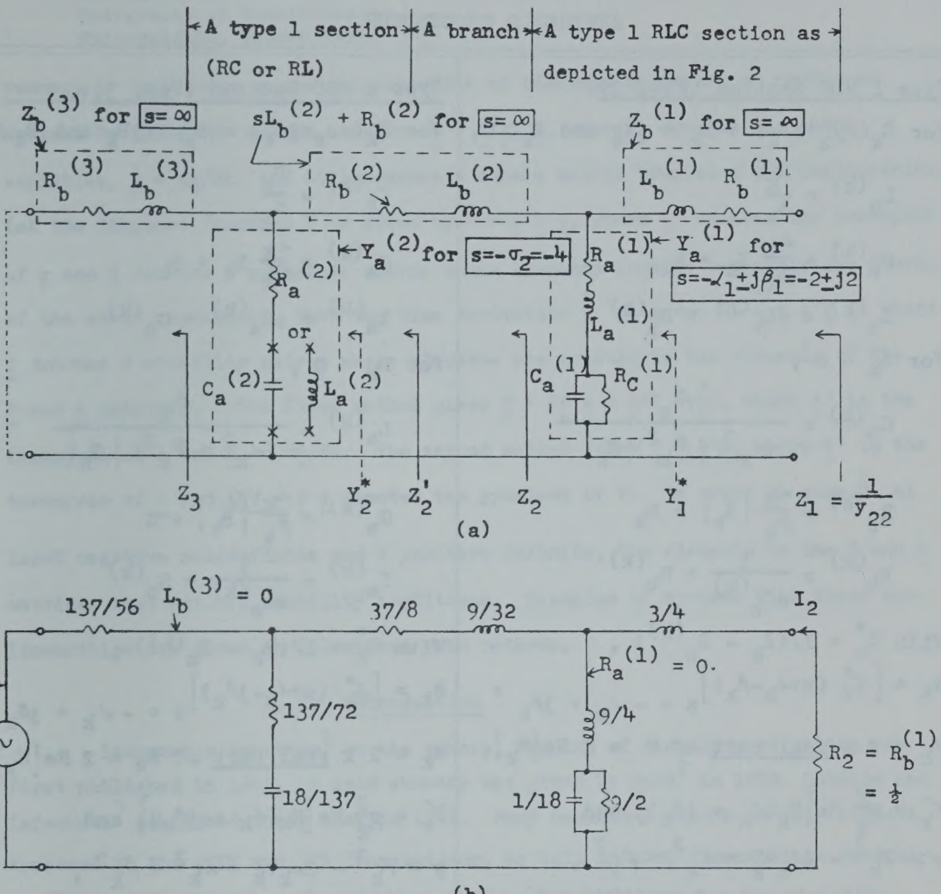

A type 1 section (RC or RL)
A branch
A type 1 RLC section as depicted in Fig. 2
Z_b^(3) for s= ∞
R_b^(3) L_b^(3)
sL_b^(2) + R_b^(2) for s= ∞
R_b^(2) L_b^(2)
Z_b^(1) for s= ∞
L_b^(1) R_b^(1)
Y_a^(2) for s=-σ_2=-4
R_a^(2)
or
C_a^(2)
L_a^(2)
R_a^(1)
Y_a^(1) for s=-α_1±jβ_1=-2±j2
L_a^(1)
C_a^(1)
R_C^(1)
Z_3
Y_2^*
Z_2'
Z_2
Y_1^*
Z_1 = 1/y_22
(a)
137/56
L_b^(3) = 0
37/8
9/32
3/4
I_2
+
E_1
137/72
18/137
R_a^(1) = 0
9/4
1/18
9/2
R_2 = R_b^(1) = ½
(b)

Fig. 9-(a) Prediction. (b) An RLC ladder arrangement as the result of synthesis.

TABLE I
SYNTHESIS PROCEDURES

Type 1 RLC Section (Fig. 2):	Type 2 RLC Section (Fig. 3):
For $Z_k(-\alpha_k+j\beta_k) = R_k + jX_k$ and $X_k \geqslant 0$,	For $Y_k(-\alpha_k+j\beta_k) = G_k + jB_k$ and $B_k \geqslant 0$,
$L_b^{(k)} = \frac{X_k}{\beta_k}$	$C_a^{(k)} = \frac{B_k}{\beta_k}$
$R_b^{(k)} = \frac{\alpha_k}{\beta_k} X_k + R_k$	$G_a^{(k)} = \frac{\alpha_k}{\beta_k} B_k + G_k$
$Z_b^{(k)} = sL_b^{(k)} + R_b^{(k)}$	$Y_a^{(k)} = sC_a^{(k)} + G_a^{(k)}$
For $X_k < 0$,	For $B_k < 0$,
$C_b^{(k)} = \frac{\beta_k}{(\alpha_k^2 + \beta_k^2)\lvert X_k \rvert}$	$L_a^{(k)} = \frac{\beta_k}{(\alpha_k^2 + \beta_k^2)\lvert B_k \rvert}$
$R_b^{(k)} = \frac{\alpha_k}{\beta_k} \lvert X_k \rvert + R_k$	$G_a^{(k)} = \frac{\alpha_k}{\beta_k} \lvert B_k \rvert + G$
$Z_b^{(k)} = \frac{1}{sC_b^{(k)}} + R_b^{(k)}$	$Y_a^{(k)} = \frac{1}{sL_a^{(k)}} + G_a^{(k)}$
With $Y_k^* = 1/(Z_k - Z_b^{(k)})$,	With $Z_k^* = 1/(Y_k - Y_a^{(k)})$,
$N_k = \left[Y_k^* \cdot (s+\alpha_k-\beta_k)\right]_{s = -\alpha_k + j\beta_k}$,	$N_k = \left[Z_k^* \cdot (s+\alpha_k-j\beta_k)\right]_{s = -\alpha_k + j\beta_k}$,
$M_k = 2 \times$ real part of $N. = 2\,\mathrm{Re}[N_k]$,	$M_k = 2 \times$ real part of $N_k = 2\,\mathrm{Re}[N_k]$,
$\gamma_k = \frac{2}{M_k} \mathrm{Re}\left[N_k \cdot (\alpha_k + j\beta_k)\right]$ and	$\gamma_k = \frac{2}{M_k} \mathrm{Re}\left[N_k (\alpha_k + j\beta_k)\right]$ and
$\varphi = \gamma_k^2 - 2\alpha_k\gamma_k + \alpha_k^2 + \beta_k^2$,	$\varphi = \gamma_k^2 - 2\alpha_k\gamma_k + \alpha_k^2 + \beta_k^2$,
$L_a^{(k)} = \frac{1}{M_k}$; $R_a^{(k)} = \frac{2\alpha_k - \gamma_k}{M_k}$	$C_b^{(k)} = \frac{1}{M_k}$; $R_b^{(k)} = \frac{M_k}{2\alpha_k - \gamma_k}$
$C_a^{(k)} = \frac{M_k}{\varphi}$; $R_C^{(k)} = \frac{\varphi}{M_k\gamma_k}$	$L_b^{(k)} = \frac{M_k}{\varphi}$; $R_L^{(k)} = \frac{M_k\gamma_k}{\varphi}$

FORMULATION OF LIAPUNOV FUNCTIONS OF NONLINEAR SYSTEMS FOR STABILITY STUDIES

Y. H. KU
The Moore School of Electrical Engineering
University of Pennsylvania
Philadelphia, Pennsylvania

Summary - This paper gives a formulation of Liapunov functions of nonlinear systems by using the matrix relation $\dot{\underline{x}} = \underline{A}(\underline{x})\ \underline{x}$, where $\underline{x}$ denotes the state variables, $\dot{\underline{x}} = d\underline{x}/dt$, and $\underline{A}(\underline{x})$ denotes a square matrix involving the nonlinearities. Let the Liapunov function V be given by $V = \underline{x}'\underline{S}\ \underline{x}$, where $\underline{x}'$ denotes the transpose of $\underline{x}$ and $\underline{S}$ denotes a symmetric matrix whose elements contain constants and functions of the state variables. Then the time derivative $\dot{V}$ is given by $\dot{V} = \underline{x}'\underline{T}\ \underline{x}$, where $\underline{T}$ denotes a stability matrix whose elements are related to the elements of the $\underline{S}$ and $\underline{A}$ matrices. The first method gives $\underline{T} = (\underline{S}\ \underline{A} + \underline{A}'\underline{S} + \dot{\underline{S}})$, where $\underline{A}'$ is the transpose of $\underline{A}$ and $\dot{\underline{S}} = d\underline{S}/dt$. The second method gives $\underline{T} = \underline{B}'\underline{A}$, where $\underline{B}'$ is the transpose of $\underline{B}$ and $\underline{\nabla V} = \underline{B}\ \underline{x}$ denotes the gradient of V. In order to make $\dot{V}$ at least negative semidefinite and V positive definite, the elements in the $\underline{T}$ and $\underline{S}$ matrices must satisfy stability conditions. Examples of systems with three nonlinearities are given to illustrate both methods.

1. Introduction

Liapunov's treatise[1] on the general problem of stability of motion was first published in 1893. A good summary was given by Hahn[2] in 1959. LaSalle and Lefschetz[3] published their book in 1961. Many technical papers on this subject appeared in the AIEE and IRE Transactions, as well as ASME Transactions recently.

This paper presents a formulation of Liapunov functions for the stability study of nonlinear systems which can be characterized by the matrix relation $\dot{\underline{x}} = \underline{A}(\underline{x})\ \underline{x}$, where $\underline{x}$ denotes the state variables in the n-dimensional phase space, $\dot{\underline{x}} = d\underline{x}/dt$, and $\underline{A}(\underline{x})$ denotes a square matrix involving the nonlinearities. The Liapunov function is given by $V = \underline{x}'\underline{S}\ \underline{x}$, where $\underline{x}'$ is the transpose of $\underline{x}$ and

---

* This work was supported by the National Science Foundation, Washington D. C.

+ The Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa.

[1] A. M. Liapunov, "Problème général de la stabilité du mouvement," Communications de la Société Mathématique de Kharkow, 1893. French translation, 1907; American edition, Princeton Univ. Press, Princeton, N. J., 1949.

[2] W. Hahn, "Theorie und Anwendung der direkten Methode von Liapunov, Ergebnisse der Mathematik und ihrer Grenzebiete," Springer-Verlag, Berlin, 1959.

[3] J. P. LaSalle and S. Lefschetz, "Stability by Liapunov's Direct Method with Applications," Academic Press, New York, 1961.

$\underline{S}$ is a symmetric matrix whose elements β_{ij} involve functions of state variables as well as constants k_{ij}. The time derivative $\dot{V}$ is given by $\dot{V} = \underline{x}'\underline{T}\,\underline{x}$, where $\underline{T}$ is a square matrix whose elements α_{ij} are related to the elements β_{ij} as well as the elements of the $\underline{A}$ matrix. Two methods are given for determining $\underline{T}$. In order to make V positive definite and $\dot{V}$ at least negative semidefinite, the elements in the $\underline{S}$ and $\underline{T}$ matrices must satisfy stability conditions. Both methods are illustrated by stability study of nonlinear systems with three nonlinearities.

2. Characterization of the Nonlinear System

Let the nonlinear third order system be represented by

$$x^{(3)} + \phi(x,\dot{x})\,x^{(2)} + g(\dot{x}) + f(x) = 0 \qquad (1)$$

where $f(x)$, $g(\dot{x})$ and $\phi(x,\dot{x})$ are nonlinear functions of x, $\dot{x}$, x and $\dot{x}$ respectively. Let $x = x_1$, $\dot{x} = dx/dt = x_2$, $x^{(2)} = d^2x/dt^2 = x_3$, and $x^{(3)} = d^3x/dt^3 = \dot{x}_3$. Then equation 1 is equivalent to the following system of first order equations:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= -f(x_1) - g(x_2) - \phi(x_1,x_2)\,x_3 \end{aligned} \qquad (2)$$

In matrix form equations 2 become

$$\dot{\underline{x}} = \underline{A}(\underline{x})\,\underline{x} \qquad (3)$$

where

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \underline{A}(\underline{x}) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \dfrac{-f(x_1)}{x_1} & \dfrac{-g(x_2)}{x_2} & -\phi(x_1,x_2) \end{bmatrix} \qquad (4)$$

3. Formulation of the Liapunov Function

The Liapunov function V can be formulated by the matrix equation

$$V = \underline{x}'\,\underline{S}\,\underline{x} > 0 \qquad (5)$$

where $\underline{S}$ is a symmetric matrix with elements β_{ij} involving functions of the state variables and the constants k_{ij}. For a third order system, $\underline{S}$ is given by

$$\underline{S} = \begin{bmatrix} k_{11} + \dfrac{\gamma_{11}+\gamma_{12}}{x_1^2} & k_{12} + \dfrac{f_{12}}{x_1} & k_{13} \\ k_{12} + \dfrac{f_{12}}{x_1} & k_{22} + \dfrac{\gamma_{22}+\gamma_{12}}{x_2^2} & k_{23} \\ k_{13} & k_{23} & k_{33} \end{bmatrix} \qquad (6)$$

where f_{12} is a function of x_1, Y_{12} is related to an integral involving $\emptyset(x_1,x_2)$,

$$Y_{11} = \int_0^{x_1} y_1(x_1)\, dx_1, \qquad Y_{22} = \int_0^{x_2} y_2(x_2)\, dx_2.$$

Substituting equation 6 into equation 5 gives

$$V = k_{11}x_1^2 + 2k_{12}x_1x_2 + 2k_{13}x_1x_3 + Y_{11} + 2Y_{12} + Y_{22} + 2f_{12}x_2$$

$$+k_{22}x_2^2 + k_{33}x_3^2 + 2k_{23}x_2x_3 \qquad (7)$$

The constants k_{ij} and the unknown functions Y_{11}, Y_{22}, Y_{12} and f_{12} are to be determined. For a fourth order system, the Liapunov function V has more terms. However, V may be said to take the general quadratic form and contain terms involving the given nonlinearities of the system under stability study.

4. First Method of Determining T.

The time derivative $\dot{V} = dV/dt$ is given by the matrix equation

$$\dot{V} = \underline{x}'\underline{T}\,\underline{x} \leq 0 \qquad (8)$$

where $\underline{T}$ is a square matrix which may be termed the stability matrix and has the elements α_{ij} which involve nonlinearities.

First Method of determining T:

$$\dot{V} = \frac{dV}{dt} = \frac{d}{dt}(\underline{x}'\underline{S}\,\underline{x}) = \dot{\underline{x}}'\underline{S}\,\underline{x} + \underline{x}'\underline{S}\,\dot{\underline{x}} + \underline{x}'\dot{\underline{S}}\,\underline{x} \qquad (9)$$

Taking the transpose of equation 3 gives

$$\dot{\underline{x}}' = \underline{x}'\,\underline{A}(\underline{x})' = \underline{x}'\,\underline{A}' \qquad (10)$$

Substituting equations 3 and 10 into equation 9 and collecting terms give

$$\dot{V} = \underline{x}'(\underline{A}'\underline{S} + \underline{S}\,\underline{A} + \dot{\underline{S}})\,\underline{x} = \underline{x}'\underline{T}\,\underline{x} \qquad (11)$$

Hence

$$\underline{T} = (\underline{A}'\underline{S} + \underline{S}\,\underline{A} + \dot{\underline{S}}) \qquad (12)$$

Since $\underline{S}$ is given as a symmetric matrix, $\underline{S} = \underline{S}'$, its transpose, and $\underline{A}'\underline{S}' = \underline{A}'\underline{S}$ is the transpose of $\underline{S}\,\underline{A}$. The sum of $(\underline{A}'\underline{S}+\underline{S}\,\underline{A})$ is symmetric, and so is $\dot{\underline{S}}$. It follow that $\underline{T}$ is a symmetric matrix.

However, so far as the contribution to $\dot{V}$ is concerned (see equation 8), the sum of $(\underline{S}\,\underline{A} + \underline{A}'\underline{S})$ can be replaced by $2(\underline{S}\,\underline{A})$. Thus equation 12 can be replaced by its equivalent form

$$\underline{T}_e = (2\,\underline{S}\,\underline{A} + \dot{\underline{S}}) \qquad (13)$$

Taking the matrix multiplication of equations 6 and 4 gives

$$2(\underline{S}\,\underline{A}) = 2\begin{bmatrix} -k_{13}\frac{f}{x_1} & k_{11}+\frac{Y_{11}+Y_{12}}{x_1^2} - k_{13}\frac{g}{x_2} & k_{12}+\frac{f_{12}}{x_1} - k_{13}\emptyset \\ -k_{23}\frac{f}{x_1} & k_{12}+\frac{f_{12}}{x_1} - k_{23}\frac{g}{x_2} & k_{22}+\frac{Y_{22}+Y_{12}}{x_2^2} - k_{23}\emptyset \\ -k_{33}\frac{f}{x_1} & k_{13} - k_{33}\frac{g}{x_2} & k_{23} - k_{33}\emptyset \end{bmatrix} \quad (14)$$

Equation 14 can be much simplified if one sets $k_{11}=k_{22}=k_{12}=k_{13}=0$, $k_{33}=1$, $k_{23}=a_3$, a positive constant, and $f_{12}(x_1)=f(x_1)$. Note that in the $\dot{\underline{S}}$ matrix (see equation 20 to follow), there are no terms involving k_{ij}. Hence from the general relation $\alpha_{ij}+\alpha_{ji}=0$ in the final $\underline{T}$ matrix, there is no term in α_{31} to balance or cancel with the k_{12} and $k_{13}\emptyset$ terms in α_{13} obtained from equation 14. Setting k_{12} and k_{13} to zero satisfies the condition $\alpha_{13}+\alpha_{31}=0$. Setting $k_{33}=1$ and $f_{12}=f(x_1)$ satisfies the same condition between the elements α_{13} and α_{31}. From the condition $\alpha_{23}+\alpha_{32}=0$, since $k_{13}=0$, k_{22} has to be zero also. Similarly, from the condition $\alpha_{12}+\alpha_{21}=0$, there are no terms in α_{21} to balance or cancel with k_{11} and k_{13}, and hence it is necessary to set $k_{11}=0$ and $k_{13}=0$. So far it is only logical to compare the terms involving k_{ij}'s and determine their possible values (0, 1, or some positive value like $k_{23}=a_3$). After simplification equation 14 becomes

$$2(\underline{S}\,\underline{A}) = 2\begin{bmatrix} 0 & \frac{Y_{11}+Y_{12}}{x_1^2} & \frac{f}{x_1} \\ -a_3\frac{f}{x_1} & \frac{f}{x_1} - a_3\frac{g}{x_2} & \frac{Y_{22}+Y_{12}}{x_2^2} - a_3\emptyset \\ -\frac{f}{x_1} & -\frac{g}{x_2} & a_3-\emptyset \end{bmatrix} \quad (15)$$

Note that elements 13 and 31 in the matrix given by equation 15 cancel with each other and may be replaced by two 0's so far as the effective contribution to $\dot{V}$ is concerned.

The $\dot{\underline{S}}$ matrix is obtained by replacing β_{ij} in the $\underline{S}$ matrix by $\dot{\beta}_{ij}$. Note that

$$\frac{d}{dt}\left(\frac{f}{x_1}\right) = \left(\frac{f'}{x_1} - \frac{f}{x_1^2}\right) x_2 \quad (16)$$

where $f' = \partial f/\partial x_1$. Similarly,

$$-\frac{d}{dt}\left(\frac{Y_{11}+Y_{12}}{x_1^2}\right) = \frac{Y'_{11}x_2+\dot{Y}_{12}}{x_1^2} - 2\left(\frac{Y_{11}+Y_{12}}{x_1^3}\right) x_2 \qquad (17)$$

$$-\frac{d}{dt}\left(\frac{Y_{22}+Y_{12}}{x_2^2}\right) = \frac{Y'_{22}x_3+\dot{Y}_{12}}{x_2^2} - 2\left(\frac{Y_{22}+Y_{12}}{x_2^3}\right) x_3 \qquad (18)$$

where $Y'_{11} = \partial Y_{11}/\partial x_1 = y_1$, $Y'_{22} = \partial Y_{22}/\partial x_2 = y_2$, and, for $Y = Y_{12}$,

$$\dot{Y}_{12} = -\frac{\partial Y}{\partial x_1} x_2 + \frac{\partial Y}{\partial x_2} x_3 \qquad (19)$$

Substituting these values in the $\dot{\underline{S}}$ matrix and rearranging terms give

$$\dot{\underline{S}} = 2 \begin{array}{|c|c|c|} \hline 0 & \frac{1}{x_1}\frac{\partial Y}{\partial x_1} - \left(\frac{Y_{11}+Y_{12}}{x_1^2}\right) & 0 \\ \hline \frac{a_3 f}{x_1} & \frac{-f}{x_1} + f' & -\left(\frac{Y_{22}+Y_{12}}{x_2^2}\right) + a_3\phi \\ \hline 0 & \frac{g}{x_2} & 0 \\ \hline \end{array} \qquad (20)$$

Adding equations 15 and 20 gives

$$\underline{T} = 2 \begin{array}{|c|c|c|} \hline 0 & \frac{1}{x_1}\frac{\partial Y}{\partial x_1} & 0 \\ \hline 0 & -\left(a_3\frac{g}{x_2} - f'\right) & 0 \\ \hline 0 & 0 & -(\phi - a_3) \\ \hline \end{array} \qquad (21)$$

It is seen from the second method that

$$y_1 = 2a_3 f(x_1), \qquad y_2 = 2g(x_2) \qquad (22)$$

and

$$Y_{12} = a_3\int_0^{x_2} \phi(x_1,x_2)\, x_2\, dx_2 \qquad (23)$$

Then equation 6 simplifies to

$$\underline{S} = \begin{array}{|c|c|c|} \hline \frac{Y_{11}+Y_{12}}{x_1^2} & \frac{f}{x_1} & 0 \\ \hline \frac{f}{x_1} & \frac{Y_{22}+Y_{12}}{x_2^2} & a_3 \\ \hline 0 & a_3 & 1 \\ \hline \end{array} \qquad (24)$$

where $Y_{11} = 2a_3\int_0^{x_1} f(x_1)dx_1 = 2a_3F(x_1)$ and $Y_{22} = 2\int_0^{x_2} g(x_2)\,dx_2 = 2G(x_2)$.

5. Second Method of Determining T.

The second method is based on the relation

$$\dot{V} = \underline{\nabla V}'\underline{x} = (\underline{x}'\underline{B}')(\underline{A}\,\underline{x}) = \underline{x}'\underline{T}\,\underline{x} \tag{25}$$

where the gradient is defined by

$$\underline{\nabla V} = \underline{B}\,\underline{x} \tag{26}$$

and the stability matrix is defined by

$$\underline{T} = \underline{B}'\underline{A} \tag{27}$$

Referring to equations 6 and 7 for a third order nonlinear system the $\underline{B}$ matrix is given by

$$\underline{B} = 2\begin{bmatrix} k_{11} + \frac{y_1}{2x_1} + \frac{1}{x_1}\frac{\partial Y}{\partial x_1} & k_{12} + f'_{12} & k_{13} \\ k_{12} + \frac{f_{12}}{x_1} & k_{22} + \frac{y_2}{2x_2} + \frac{1}{x_2}\frac{\partial Y}{\partial x_2} & k_{23} \\ k_{13} & k_{23} & k_{33} \end{bmatrix} \tag{28}$$

Note that for a linear system with constant coefficients, $\underline{B} = 2\underline{S}$, where $\beta_{ij} = k_{ij}$. From equation 27, the $\underline{T}$ matrix is given by

$$\underline{T} = 2\begin{bmatrix} -k_{13}\frac{f}{x_1} & k_{11} + \frac{y_1}{2x_1} + \frac{1}{x_1}\frac{\partial Y}{\partial x_1} - k_{13}\frac{g}{x_2} & k_{12} + \frac{f_{12}}{x_1} - k_{13}\phi \\ -k_{23}\frac{f}{x_1} & k_{12} + f'_{12} - k_{23}\frac{g}{x_2} & k_{22} + \frac{y_2}{2x_2} + \frac{1}{x_2}\frac{\partial Y}{\partial x_2} - k_{23}\phi \\ -k_{33}\frac{f}{x_1} & k_{13} - k_{33}\frac{g}{x_2} & k_{23} - k_{33}\phi \end{bmatrix} \tag{29}$$

where $f'_{12} = \partial f_{12}/\partial x_1$. From the general relations among the elements of the $\underline{T}$ matrix:

$$\alpha_{jj} \leq 0, \qquad \alpha_{ij} + \alpha_{ji} = 0 \tag{30}$$

the constants are set as follows: $k_{11} = k_{22} = k_{12} = k_{13} = 0$, $k_{33} = 1$, and $k_{23} = a_3 > 0$. Then from the relation $\alpha_{13} + \alpha_{31} = 0$ it is obtained

$$f_{12}(x_1) = f(x_1) \tag{31}$$

The condition $\alpha_{12}+\alpha_{21}=0$ gives, for $k_{11}=k_{13}=0$ and $k_{23}=a_3$,

$$y_1 = 2a_3 f(x_1) \tag{32}$$

The condition $\alpha_{23}+\alpha_{32}=0$ gives, for $k_{22}=k_{13}=0$ and $k_{23}=a_3$, $k_{33}=1$,

$$y_2 = 2\,g(x_2) \tag{33}$$

and

$$Y = Y_{12} = a_3\int_0^{x_2} \phi(x_1,x_2)\, x_2\, dx_2 \tag{34}$$

Substituting these values into equation 29 and cancelling out terms (with opposite signs in elements α_{ij} and α_{ji}) give the effective $\underline{T}$ matrix:

$$\underline{T} = 2\begin{bmatrix} 0 & \frac{1}{x_1}\frac{\partial Y}{\partial x_1} & 0 \\ 0 & -(a_3-\frac{g}{x_2}-f') & 0 \\ 0 & 0 & -(\phi-a_3) \end{bmatrix} \tag{35}$$

The conditions of stability are:

$$a_3 > 0, \qquad \frac{g(x_2)}{x_2} > 0 \tag{36}$$

$$\Delta_2 = (a_3-\frac{g}{x_2}-f') > 0 \tag{37}$$

$$\Delta_3 = (\phi - a_3) > 0 \tag{38}$$

$$\Delta_2\, x_2^2 + \Delta_3\, x_3^2 - a_3 x_2\int_0^{x_2} \frac{\partial\phi}{\partial x_1}\, x_2\, dx_2 > 0 \tag{39}$$

By equation 25

$$\frac{\dot{V}}{2} = -\left[a_3\frac{g(x_2)}{x_2} - f'(x_1)\right] x_2^2 - \left[\phi(x_1,x_2)-a_3\right]x_3^2 + a_3 x_2\int_0^{x_2} \frac{\partial\phi}{\partial x_1} x_2 dx_2 \tag{40}$$

By equation 5

$$\frac{V}{2} = a_3 F(x_1) + G(x_2) + f(x_1)x_2 + a_3\int_0^{x_2} \phi(x_1,x_2)x_2 dx_2 + \frac{1}{2}x_3^2 + a_3 x_2 x_3 \tag{41}$$

If ϕ is substituted for a_3 in equation 41, there is obtained

$$\frac{V}{2} = a_3 F(x_1) + G(x_2) + f(x_1)x_2 + \frac{1}{2}(a_3 x_2 + x_3)^2 \tag{42}$$

6. Liapunov Function of Fourth Order Nonlinear System.

Let the fourth order nonlinear system be represented by

$$x^{(4)} + a_4 x^{(3)} + \phi(x,\dot{x})\, x^{(2)} + g(\dot{x}) + f(x) = 0 \tag{43}$$

Equation 3 still holds with

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \qquad \underline{A}(\underline{x}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{-f(x_1)}{x_1} & \frac{-g(x_2)}{x_2} & -\phi(x_1,x_2) & -a_4 \end{bmatrix} \tag{44}$$

where $x_1 = x$, $x_2 = \dot{x}$, $x_3 = d^2x/dt^2 = x^{(2)}$, $x_4 = d^3x/dt^3 = x^{(3)}$, and $\dot{x}_4 = x^{(4)}$.

The Liapunov function is given by equation 5, where

$$\underline{S} = \begin{bmatrix} \frac{Y_{11}+Y_{12}}{x_1^2} & a_4 \frac{f}{x_1} & \frac{f}{x_1} & 0 \\ a_4 \frac{f}{x_1} & \frac{Y_{22}+Y_{12}}{x_2^2} & k_{23}+\frac{g}{x_2} & k_{24} \\ \frac{f}{x_1} & k_{23}+\frac{g}{x_2} & \phi & k_{34} \\ 0 & k_{24} & k_{34} & 1 \end{bmatrix} \tag{45}$$

Note that the constants k_{ij} are set as follows: $k_{11}=k_{22}=k_{33}=k_{12}=k_{13}=k_{14}=0$, $k_{44}=1$. The functions f_{12} and f_{13} are set as $f_{12}(x_1) = a_4 f(x_1)$ and $f_{13}(x_1)=f(x_1)$, respectively. Similarly, $g_{23}(x_2)$ is set as equal to $g(x_2)$. Then equation 5 gives

$$V = Y_{11} + Y_{22} + 2Y_{12} + 2a_4 f(x_1)x_2 + 2f(x_1)x_3 + 2k_{23}x_2x_3 + 2g(x_2)x_3$$
$$+2k_{24}x_2x_4 + \phi(x_1,x_2)\, x_3^{\,2} + x_4^{\,2} + 2k_{34}x_3x_4 \tag{46}$$

where $k_{34}=a_4$, $k_{24}=a_4^{\,2}$ and $k_{23}=a_4^{\,3}$ are to be determined. Again Y_{11} is related to the integral of $y_1(x_1)$, Y_{22} is related to the integral of $y_2(x_2)$, and Y_{12} is related to an integral involving $\phi(x_1,x_2)$.

By First Method of Determining $\underline{T}$:

$$2\,\underline{S}\,\underline{A} = 2\begin{bmatrix} 0 & \frac{Y_{11}+Y_{12}}{x_1^2} & 0 & 0 \\ \frac{-a^2 f}{x_1} & \frac{af}{x_1} - \frac{a^2 g}{x_2} & \frac{Y_{22}+Y_{12}}{x_2^2} - a^2\phi & 0 \\ 0 & \frac{f}{x_1} - \frac{ag}{x_2} & a^3 + \frac{g}{x_2} - a\,\phi & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{47}$$

where $a = a_4$. Differentiating β_{ij} in equation 45 with respect to time gives

$$\dot{\underline{S}} = 2 \begin{bmatrix} 0 & \frac{1}{x_1}\frac{\partial Y}{\partial x_1} - \frac{\left(\frac{Y_{11}+Y_{12}}{2}\right)}{x_1} & 0 & 0 \\ \frac{a^2 f}{x_1} & \frac{-af}{x_1} + af' & -\frac{\left(\frac{Y_{22}+Y_{12}}{2}\right)}{x_2} + a^2\phi & 0 \\ 0 & \frac{-f}{x_1} + \frac{ag}{x_2} & g' - \frac{g}{x_2} + \frac{\dot{\phi}}{2} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{48}$$

Then

$$\underline{T} = (2\,\underline{S}\,\underline{A} + \dot{\underline{S}}) = 2 \begin{bmatrix} 0 & \frac{1}{x_1}\frac{\partial Y}{\partial x_1} & 0 & 0 \\ 0 & -a\left(\frac{ag}{x_2} - f'\right) & 0 & 0 \\ 0 & 0 & -a\phi + \frac{\dot{\phi}}{2} + g' + a^3 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{49}$$

where $f' = \partial f/\partial x_1$, $g' = \partial g/\partial x_2$, $\dot{\phi} = (\partial\phi/\partial x_1)x_2 + (\partial\phi/\partial x_2)\,x_3$, and

$$Y = Y_{12} = \int_0^{x_2} (a^2\phi - f')\,x_2\,dx_2 \tag{50}$$

Then the time derivative $\dot{V}$ is given by

$$\frac{\dot{V}}{2} = -a\left(\frac{ag}{x_2} - f'\right)x_2^2 - \left(a\phi - \frac{\dot{\phi}}{2} - g' - a^3\right)x_3^2 + x_2\int_0^{x_2}\left(a^2\frac{\partial\phi}{\partial x_1} - f''\right)x_2\,dx_2 \tag{51}$$

By Second Method of Determining $\underline{T}$:

$$\underline{B} = 2 \begin{bmatrix} \frac{y_1}{2x_1} + \frac{1}{x_1}\frac{\partial Y}{\partial x_1} & af' & f' + \frac{x_3}{2}\frac{\partial\phi}{\partial x_1} & 0 \\ \frac{af}{x_1} & \frac{y_2}{2x_2} + \frac{1}{x_2}\frac{\partial Y}{\partial x_2} & k_{23} + g' + \frac{x_3}{2}\frac{\partial\phi}{\partial x_2} & k_{24} \\ \frac{f}{x_1} & k_{23} + \frac{g}{x_2} & \phi & k_{34} \\ 0 & k_{24} & k_{34} & 1 \end{bmatrix} \tag{52}$$

Then $\underline{T} = \underline{B}'\underline{A}$ gives

$$\underline{T} = 2 \left|\begin{array}{c|c|c|c} 0 & \frac{y_1}{2x_1} + \frac{1}{x_1}\frac{\partial Y}{\partial x_1} & \frac{af}{x_1} & \frac{f}{x_1} \\ \hline \frac{-k_{24}f}{x_1} & af' - k_{24}\frac{g}{x_2} & \frac{y_2}{2x_2} + \frac{1}{x_2}\frac{\partial Y}{\partial x_2} - k_{24}\phi & k_{23} + \frac{g}{x_2} - k_{24}a \\ \hline \frac{-k_{34}f}{x_1} & f' + \frac{x_3}{2}\frac{\partial \phi}{\partial x_1} - k_{34}\frac{g}{x_2} & k_{23} + g' + \frac{x_3}{2}\frac{\partial \phi}{\partial x_2} - k_{34}\phi & \phi - k_{34}a \\ \hline \frac{-f}{x_1} & \frac{-g}{x_2} & k_{24} - \phi & k_{34} - a \end{array}\right| \tag{53}$$

From the general relations $\alpha_{jj} \leq 0$ and $\alpha_{ij} + \alpha_{ji} = 0$, there are obtained:

$$k_{34} = a, \quad k_{24} = a^2, \quad k_{23} = a^3 \tag{54}$$

$$y_1 = 2a^2\, f(x_1), \quad y_2 = 2a\, g(x_2), \quad Y = Y_{12} = \int_0^{x_2} (a^2\phi - f')\, x_2 dx_2 \tag{55}$$

The simplified $\underline{T}$ matrix is then given by

$$\underline{T}_e = 2 \left|\begin{array}{c|c|c|c} 0 & \frac{1}{x_1}\frac{\partial Y}{\partial x_1} & 0 & 0 \\ \hline 0 & -a(\frac{ag}{x_2} - f') & 0 & 0 \\ \hline 0 & 0 & -a\phi + (\dot{\phi}/2) + g' + a^3 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right| \tag{56}$$

The conditions of stability are:

$$a_4 > 0, \quad \frac{g(x_2)}{x_2} > 0, \quad (\frac{a_4 g}{x_2} - f') > 0 \tag{57}$$

$$(a_4\phi - \frac{\dot{\phi}}{2} - g' - a_4^3) > 0 \tag{58}$$

$$(a_4^2 \frac{\partial \phi}{\partial x_1} - f'')\, x_2 \leq 0 \tag{59}$$

The final form of the Liapunov function V is given by

$$\begin{aligned} \frac{V}{2} = {} & a_4^2 F(x_1) + a_4 G(x_2) + a_4 f(x_1) x_2 + f(x_1) x_3 + g(x_2) x_3 \\ & + a_4^3 x_2 x_3 + a_4^2 x_2 x_4 + a_4 x_3 x_4 + \frac{1}{2}\phi(x_1, x_2)\, x_3^2 + \frac{1}{2}\, x_4^2 \\ & + \int_0^{x_2} \left[a_4^2 \phi(x_1, x_2) - f'(x_1) \right] x_2\, dx_2 \end{aligned} \tag{60}$$

where $f'' = \partial^2 f/\partial x_1^2$, $F(x_1) = \int_0^{x_1} f(x_1) dx_1$, and $G(x_2) = \int_0^{x_2} g(x_2)\, dx_2$.

7. Concluding Remarks

Two methods are presented for determining a square matrix $\underline{T}$ which gives stability conditions according to $\dot{V} = \underline{x}'\underline{T}\,\underline{x} \leq 0$. The constants and unknown functions determined give the elements β_{ij} in a symmetric matrix $\underline{S}$ which in turn gives $V = \underline{x}'\underline{S}\,\underline{x} > 0$. The method can be extended to high order systems with three or more non-linearities and time-varying coefficients.

8. Acknowledgments

The author is deeply indebted to the published works of many authors who contributed to the understanding and significance of Liapunov's direct method. He wishes especially to acknowledge the support of the National Science Foundation, Washington D. C. and the constant encouragement of Drs. J. G. Brainerd and C. N. Weygandt of the Moore School of E. E., University of Pennsylvania. He wishes to thank Dr. S. D. Bedrosian for presenting the paper for him at the First Allerton Conference on Circuit and System Theory.

NONLINEAR SYSTEM MODELS

THEORDORE A. BICKART
Department of Electrical Engineering
Syracuse University
Syracuse, New York

ABSTRACT

The concept of a system, nonlinear or linear, as a transformation in signal space, where signals are vectors, is developed. Guides to the creation of system models are set forth as they evolve from signal vector representations and from functional representations of the signal space transformation.

Introduction

The creation of a model - a representation - of a system is the first step in achieving an analysis of that system. A number of representations have been presented in the literature; the references cited at the end of this paper discuss some of the notable contributions. Certainly, these are not the only representations, nor, is it anticipated, the best. One could concoct other representations, probably without a guide to provide a sense of direction in the quest to establish a representation.

In this paper we will develop an idea which leads to a new concept of a system. This concept will lead us to some general statements about system representations. These statements will serve as a suitable guide as one goes about the task of putting a model together. We will begin to put our thoughts together by considering signal representations.

A signal is an observable which is conceivable when given a representation. One representation might be the time function which indicates the strength of the signal at each instant of time. Another representation might be the frequency function which indicates the strength of the signal for every value of frequency.

The concept of a signal as an observable makes it quite independent of the representation by which we conceive of the signal. Let us then adopt the idea of a signal as being a vector in signal space. This is a consistent idea, in that we associate the picture of an object - a vector - fixed in space with that unperturbable observable - a signal.

Our signal vector may acquire a representation only after we establish a basis for the signal space and a norm by which to measure the strength of the signal vector. The basis provides a set of fixed directions in a signal space and, hence, the means of specifying the direction of the signal vector (relative to the signal space basis).[1]

A system is the link between signals. Given one or more input (or stimulus) signals the system yields one or more output (or response) signals. For the sake of convenience, we will restrict our discussion to systems having a single stimulus signal and a single response signal.

The concept of a system is that of an observable which is a transformation in signal space. Quite apart from the representation of a system, we may think of it as being the transformation which maps our signal space into itself. A representation for a system will depend, therefore, upon the basis, by which orientation is provided, in signal space.

Let us imagine, now, that our system is a vector in system space. System space is to be the space of all transformations which map signal space into itself. The object of invoking this image of a system is to draw a parallel with the vector image of a signal. We see, then, that the representation of a system will depend upon the basis of the system space and upon the norm by which the system vector is measured in system space.

We will consider, first of all, signal representations, since the concomitant establishment of a basis for signal space must precede consideration of system representations.

Signal Representations

Let us start by considering a signal that is known to us because we know its time function, s(t); that is, at any instant in time, t, our signal has the strength s(t). The basis of our signal space is the sampling function at that instant in time; the projection of the signal vector on the basis vector is s(t).

On our quest for system representations, we will find it expedient to consider system representations which stem from signal representations on an interval of time. Several intervals of time are of greatest utility. There is the semi-infinite time interval called the past and present. Let us note that the present time is not fixed but is changing relative to that epoch which serves as our time datum. We will return to this point later. There is, also, the semi-infinite time interval which we call the present and future and the bi-infinite time interval which is the past, present, and future. We might also want to consider the semi-infinite intervals which are the time datum and all prior time or all subsequent time or the bi-infinite interval which is the time datum and all prior and subsequent time.

Let us now turn our attention to what we shall call a discrete basis for signal space. Consider the discretely indexed set of functions $\phi_i(t)$ which are complete and orthonormal with respect to the weighting function w(t) over the time interval (a,b)[2]. (Note, orthogonality is not required in developing representations; it is usually desirable, though.) The $\phi_i(t)$ may serve as a basis for signal space on the time interval (a,b). If we adopt a minimum mean square error approximation requirement, then the projection of our signal vector on the basis vectors (functions), $\phi_i(t)$, may be determined according to:

$$s_i = \int_a^b s(t)\, \phi^*_i(t)\, w(t)\, dt \qquad (1)$$

The signal vector will then have as one representation:[4]

$$\sum_i s_i \phi_i(t) \tag{2}$$

The value of the signal at any instant of time, t, in the interval (a,b) is given by (2). The ordered set of numbers s_i are often called a signal vector; however, to avoid confusion we will, for the moment, call them a numerical signal vector.

Equation (1) establishes a representation of our signal vector relative to an epoch. Let us replace (1) by:

$$s_i(t) = \int_a^b s(\tau + t)\, \phi_i^*(\tau)\, w(\tau)\, d \tag{3}$$

and hence (2) by:

$$\sum_i s_i(t)\, \phi_i(\tau) \tag{4}$$

(4) gives us the value of the signal at any instant of time $\tau + t$. If t is the present time, then the present value of the signal (assuming that the interval (a,b) includes the origin) is:

$$\sum_i s_i(t)\, \phi_i(0) \tag{5}$$

(3), we see, establishes a representation relative to the present time. The numerical vector associated with our fixed signal vector is a function of time since the present time changes relative to the time datum. Note also that the orientation of the signal space basis is changing with time; it provides a representation of the signal over some time interval relative to the present.

If in a limiting process we replace the summation operation, with respect to a discrete index in (2) and (4), by an integration, with respect to a continuous index (variable), then we obtain what will be called a continuous basis for signal space. Corresponding to (1), we have:

$$s(\lambda) = \int_a^b s(t)\, \phi^*(,\lambda)\, w(t)\, dt \tag{6}$$

The signal vector representation, similar to (2) is:

$$\int s(\lambda)\, \phi(\lambda, t)\, d\lambda \tag{7}$$

The basis of our signal space is the kernal function $\phi(t,\lambda)$ and the numerical signal vector $s(\lambda)$ is obtained by (6) which is the transformation from the basis giving us $s(t)$ to the basis extablished by $\phi(t,\lambda)$. We recognize that

$s(\lambda)$ is what we normally call a transform.

The continuous representation comparable to (3) and (4) follows as:

$$s(t,\lambda) = \int_a^b s(t + \tau)\ \phi^*(\tau,\lambda)\ w(\tau)\ d \tag{8}$$

and:

$$\int s(t,\lambda)\ \phi(\lambda,\tau)\ d\lambda \tag{9}$$

Let us stop for a moment and take stock of what we have obtained. We started with the time function representation of our signal vector. Transformations have given us discrete representations of our signal vector on an interval of time. Other transformations have provided continuous representations of the signal vector on an interval of time. In each case we have used the numerical signal vector to obtain a weighted sum of the basis vectors. The result is, really, a representation of the signal vector in the time domain. Now, rather than thinking of $s(\lambda)$ as the numerical signal vector associated with the basis established by $\phi(\lambda,t)$, let us suppose that $s(\lambda)$ is a representation of the signal vector stemming from the basis which is the sampling function at the point λ in the λ domain.

The next stage in our discussion of signal representations follows by analogy to the representations in the time domain. First, it may be that we will want to consider representations which follow from signal representations on some specified region of the λ domain. Next, we anticipate that signal vector representations will be sought which are discrete representations in the λ domain and which are continuous representations in the λ domain.

We have singled out the time domain representations to serve as a point of departure in our discussion of signal vector representations. We have reaches the stage, now, at which it is expedient to phrase our comments in terms of representations in arbitrary domains.

Let us suppose that $s(\lambda_1)$ and $s(\lambda_2)$ are signal vector representations in the λ_1 and λ_2 domains respectively. We note that $s(\lambda_1)$ is also the numerical signal vector in the transformation which yields a λ_2 domain representation from the λ_1 domain representation. If the λ_1 domain and the λ_2 domain are both continuous (λ_1 and λ_2 are continuous variables) then we expect to have a λ_1 domain representation of the form:

$$\int s(\lambda_2)\ \phi(\lambda_2,\lambda_1)\ w(\lambda_2)\ d\lambda_2 \tag{10}$$

and a λ_2 domain representation of the form:

$$\int s(\lambda_1)\Psi(\lambda_1,\lambda_2)v(\lambda_1)\ d\lambda_1 \tag{11}$$

If the λ_1 domain is continuous and the λ_2 domain is discrete (λ_2 takes on discrete values), then we expect a λ_1 domain representation of the form:

$$\sum_{\lambda_2} s(\lambda_2)\phi(\lambda_2,\lambda_1)\ w(\lambda_2) \tag{13}$$

(11) is still an expected λ_2 domain representation. If both the λ_1 domain and the λ_2 domain are discrete, then in the λ_2 domain we expect a representation of the form:

$$\sum_{\lambda_1} s(\lambda_1)\Psi(\lambda_1,\lambda_2)\ v(\lambda_1) \tag{13}$$

Let us begin to draw our examination of signal vector representations to a close with a discussion of measurement. If we measure the strength (in some specified units) of our signal vector, at every point in the λ_1 domain we obtain $s(\lambda_1)$. If we measure the strength of the projection of the signal vector on a basis of signal space in the λ_1 domain according to (11) or (13), we get a representation of the signal vector in the λ_2 domain. If we treat the representation, $s(\lambda_2)$, in the λ_2 domain as a numerical vector then we find that our measurement process will yield us representations such as (10) or (12) in the λ_1 domain or some portion of the λ_1 domain.

The remaining point to be brought out is that of representations over some portion of a domain, for example, over the past and present in the time domain. Let us suppose that we want a representation of our signal vector in a region, say λ_1^*, of the λ_1 domain. We anticipate that there may exist a λ_2 domain such that the signal vector $s(\lambda_2)$ transformed according to (10) or (12) is equal to $s(\lambda_1)$ in the λ_1^* region of the λ_1 domain.

The essential point is that the signal vector in the λ_2 domain carries the information which gives the value of $s(\lambda_1)$ at every point in λ_1^*. Now, let us note that, if we consider a signal vector representation over only some region of the λ_2 domain, we cannot expect the representation in the λ_1 domain obtained via (10) or (12) to be over just the λ_1^* region or some part of it. The significance of this is that one cannot in general limit the region of representation in more than one domain. For example, one cannot represent the same signal on both a finite portion of the time domain and a finite portion of the frequency domain.

System Representations

Let us commence our discussion of system representations by considering a system which has as its representation $F\,[s(\lambda_1)]$. $F\,[\quad]$ is the functional which relates the system stimulus signal, $s(\lambda_1)$, in the λ_i domain to the system response signal, $r(\lambda_j)$, in the λ_j domain. We would expect that in most circumstances the stimulus signal and response signal would have representations in the same domain.

To every point in signal space, which defines a stimulus signal vector, there is another point, which defines the corresponding response signal vector. The system is an observable which is the transformation, in signal space, providing the connection between stimulus signal and response signal. Therefore, as a matter of convenience, let us envision this observable as a vector in system space. We see, then, that the representation of a system will, once we have adopted representations of our signals, depend upon the selected basis for system space and upon the norm we adopt.

Let us suppose that we have available to us a complete set of orthonormal functionals, $\left\{\Phi_m[s(\lambda_i)]\right\}$, over some portion of signal space.[5] Then, if we employ $\left\{\Phi_m[s(\lambda_i)]\right\}$ as the basis for system space, we anticipate a representation for our system of the following form:

$$\sum_m F_m \Phi_m[s(\lambda_i)] \tag{14}$$

where

$$F_m = \int_\sigma F[s(\lambda_i)]\, \Phi_m^*[s(\lambda_i)]\, W[s(\lambda_i)] ds(\lambda_i) \tag{15}$$

F_m is the numerical system vector associated with the basis vectors $\Phi_m[s(\lambda_i)]$.

We have a representation of our system, (14), in the signal space domain (or some portion thereof) in terms of the basis vectors, $\Phi_m[s(\lambda_i)]$, and the numerical system vector, F_m. It is interesting, by analogy to the representations of signals, to conjecture that F_m is a system vector in a domain other than signal space; we might call it ficticious signal space.

Let us continue by considering the representation of our system in two different domains. $F[\Lambda_m(\lambda_i)]$ and $F[\Lambda_n(\lambda_i)]$ are to be taken as vector representations in two domains of system space. If both domains are continuous, then in the $\Lambda_m(\lambda_i)$ domain we expect to find a system representation of the form:

$$\int F[\Lambda_n(\lambda_j)]\ \ \Phi[\Lambda_n(\lambda_j),\ \Lambda_m(\lambda_i)]\ W[\Lambda_n(\lambda_j)]\ d\Lambda_n(\lambda_j) \tag{16}$$

and, in the $\Lambda_n(\lambda_j)$ domain, a representation of the form:

$$\int F[\Lambda_m(\lambda_i)]\ \ \Psi[\Lambda_m(\lambda_i),\ \Lambda_n(\lambda_j)]\ V[\Lambda_m(\lambda_i)]\ d\Lambda_m(\lambda_i) \tag{17}$$

If either or both of the domains are discrete, then in place of (16) and/or (17) we have:

$$\sum_{\Lambda_n(\lambda_j)} F[\Lambda_n(\lambda_j)]\ \Phi[\Lambda_n(\lambda_j),\ \Lambda_m(\lambda_i)]\ W[\Lambda_n(\lambda_j)] \tag{18}$$

and/or:

$$\sum_{\Lambda_m(\lambda_i)} F[\Lambda_m(\lambda_i)]\,\Psi[\Lambda_m(\lambda_i),\ \Lambda_n(\lambda_j)]\,V[\Lambda_m(\lambda_i)] \tag{19}$$

Note that (16), (17), (18), and (19) may give representations only in some portion of their respective domains.

System Representations - Specific

Thus far, we have established that system representations may be categorized according to the adopted basis for and norm in signal space and the chosen basis for and norm in system space. This means that, for all practical purposes, there are an unlimited number of system representations. Let us see how some of those already given in the literature fit into the scheme of things.

The representation given by Wiener[6] starts with a stimulus signal representation in the past of the time domain using the Laguerre functions of the past as the basis for signal space. The multivariable Hermite polynomials form the basis for system space. If the coefficients of the numerical signal vector are s_i and if $H_m[\quad]$ denotes the m-th degree Hermite polynomial, then the Wiener representation gives us the present strength of the response signal as:

$$\sum_{m_1, m_2, \ldots} F_{m_1 m_2 \ldots}\, H_{m_1}[s_1]\, H_{m_2}[s_2] \ldots \tag{20}$$

This representation of non-anticipatory systems is one of an unlimited number which separate the system model into two pieces. In this case, the Laguerre filter network, which is linear and which generates the numerical signal vector s_i, and the Hermite network, which is a no-memory nonlinear device realizing (20). The primary advantage of the Wiener representation stems from the time average measurement which may be used to determine $F_{m_1 m_2 \ldots}$ when the stimulus signal is a Gaussian random signal.

The representation that Bose[7] gives us differs from Wiener's representation in the choice of a basis for system space. Bose defines the gate functions $G_{m_i}[s_i]$ as:

$$\begin{aligned} G_{m_i}[s_i] &= 1 \quad \ell_i + (m_i - 1)\,d_i \le s_i \le \ell_i + m_i d_i \quad m_i = 1, \ldots, M_i - 1 \\ &= 1 \quad \ell_i + (m_i - 1)\,d_i \le s_i \le u_i \qquad\quad m_i = M_i \\ &= 0 \quad \text{all other } s_i \end{aligned}$$

where, if the range of s_i is from ℓ_i to u_i, then $d_i = (u_i - \ell_i)/M_i$. Products of the gate functions create a set of multivariable gate functions which may be taken as the basis for system space. The Bose representation gives the present strength of the response signal as:

$$\sum_{m_1, m_2, \dots} F_{m_1 m_2 \dots} G_{m_1}[s_1]\, G_{m_2}[s_2] \dots \tag{21}$$

The multivariable gate functions are orthonormal, but the finite set is not complete; however, as the M_i become sufficiently large the error due to incompleteness becomes insignificant, under reasonable restrictions. A primary advantage of the multivariable gate functions is the resultant orthonormality in the signal space time domain as well as in signal space.

The Volterra series representation[8] of a system is based upon the time domain representation of the stimulus signal, with time taken relative to the present. The basis for system space is the set of functionals $H_m[\quad]$:

$$H_m[s(t), \dots, s(t)] = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} h_m(\tau_1, \dots, \tau_m) s(t - \tau_1) \dots s(t - \tau_m)\, d\tau_1 \dots d\tau_m \tag{22}$$

where the $h_m(\tau_1, \dots, \tau_m)$ are a set of kernals associated with the system. The representation for the present value of the response signal is expressed as:

$$\sum_m F_m H_m[s(t), \dots, s(t)] \tag{23}$$

The disadvantage to the Volterra series (or functional power series) is that of determining the series; the difficulty is in determining the $H_m[\quad]$, which are not, in general, orthogonal, and the F_m. Analysis may be somewhat easier using the functional power series over other representations;[9] thus, the effort required to establish (23) may be worthwhile.

If we represent our stimulus signal vector in the domain of the complex frequency associated with the Laplace transform, then the multi-dimensional transform theory developed by George[10] gives us, for the value of the response signal at a time t (relative to a zero time datum):

$$\sum_m F_m T_m[s(\varsigma), \dots, s(\varsigma)] \tag{24}$$

where:

$$T_m[s(\varsigma), \dots, s(\varsigma)] = \int_{BR} \dots \int_{br} h_m(\varsigma_1, \dots, \varsigma_m) s(\varsigma_1) \dots s(\varsigma_m)\, \epsilon^{(\varsigma_1 + \dots + \varsigma_m)t}\, d\varsigma_1 \dots d\varsigma_m \tag{25}$$

The advantage obtained in this representation is realized in analysis; interconnected sub-systems have $h_m(s_1,\ldots,s_m)$ which depends upon the $h_k(s_1,\ldots,s_k)$ of the several sub-systems.

Conclusions

The intent of this paper has been to point the way toward an unlimited number of nonlinear system representations - a number of which, it is hoped, will be found to have great utility in the analysis of nonlinear systems. We have shown how several representations, which have been devised and studied fit into the scheme evolving from our discussion. Particular representations may be difficult to work with, but they stem from conceptually simple ideas. Thus, the fact that a system model is a representation of a transformation in signal space should not be obscured by what may be a large amount of difficult work in obtaining the system representation.

Let us recapitulate the theme of this paper. A signal is a vector in signal space, comprehendable after the basis for and norm in signal space is specified. A system, which is a transformation in signal space, is a vector in system space, comprehendable after the basis for and norm in system space is specified.

References

1. Bose, A.G., "A Theory of Nonlinear Systems," MIT Research Laboratory of Electronics, TR-309, May 1956.

2. Brilliant, Martin B., "Theory of the Analysis of Nonlinear Systems," MIT Research Laboratory of Electronics, TR-345, March 1958.

3. Cameron, R.H. and Martin, W.T., "The Orthogonal Development of Non-Linear Functionals in Series of Fourier-Hermite Functionals," Annals of Mathematics, Vol. 48, No. 2, April 1947.

4. Jeutsch, Ralph, "Nonlinear Transformations of Random Processes," Prentice-Hall, Inc. (Englewood Cliffs, N.J., 1962).

5. Frank, P. and McFee, R., "Determining Input-Output Relationships of Nonlinear Systems by Inversions," IEEE Trans. on Circuit Theory, Vol. CT-10, No. 2, June 1963.

6. George, D.A., "Continuous Nonlinear Systems," MIT Research Laboratory of Electronics, TR-355, July 1959.

7. Huggins, W.H., "Are Physical Systems Abstract or Concrete?" IRE Trans. on Circuit Theory, Vol. CT-9, No. 4, December 1962.

8. Ikehara, S., "A Method of Wiener in a Nonlinear Circuit," MIT Research Laboratory of Electronics, TR-217, December 1951.

9. Lai, David C., "Representation and Analysis of Signals Part VI: Signal-Space Concepts and Dirac's Notation," John Hopkins University, Department of Electrical Engineering, AFCRC TN-60-167, January 1960.

10. Lerner, R.M., "The Representation of Signals," IRE Trans. on Circuit Theory, Vol. CT-6, Special Supplement, May 1959.

11. Singleton, H.E., "Theory of Nonlinear Transducers," MIT Research Laboratory of Electronics, TR-317, July 1956.

12. Smets, H.B., "Analysis and Synthesis of Nonlinear Systems," IRE Trans. on Circuit Theory, Vol. CT-7, No. 4, December 1960.

13. Van Trees, Harry L., "Synthesis of Optimum Nonlinear Control Systems," MIT Press (Cambridge, Massachusetts, 1962).

14. Volterra, Vito, "Theory of Functionals and of Integral and Integro-Differential Equations," Dover Publications, Inc. (New York, 1959).

15. Wiener, Norbert, "Nonlinear Problems in Random Theory," John Wiley and Sons, Inc. (New York, 1958).

16. Zadeh, L.A., "Optimum Nonlinear Filters," Journal of Applied Physics, Vol. 24, No. 4, April 1953.

17. Zames, G., "Nonlinear Operators for System Analysis," MIT Research Laboratory of Electronics, TR-370, August 1960.

Footnotes

1. See W. H. Huggins' Editorial (7) for elaboration on this concept of a signal and its representation.

2. The orthonormality condition may be expressed as:

$$\int_a^b \phi_i(t)\phi_j^*(t)w(t)dt \;=\; 1 \;,\quad i = j$$

$$= \; 0 \;,\quad i \neq j$$

3. Equation (1) is the transformation from the basis which gives us s(t) to the basis established by $\left\{\phi_i(t)\right\}$.

4. (2) is not only a representation of the signal vector but is a series expression for s(t) over the time interval (a,b).

5. The orthnormality condition may be expressed as:

$$\int_\sigma \Phi_m[s(\lambda_i)]\, \Phi_n^*[s(\lambda_i)]\, W[s(\lambda_i)]\, ds(\lambda_i)$$

where W[ ] is a weighting functional and σ denotes the portion of signal space over which $\left\{\Phi_m[s(\lambda_i)]\right\}$ is complete and orthonormal.

6. See references [1, 4, 13, 15]

7. See reference [1]

8. See references[14, 2, 6, 12, 13,]

9. See references [5, 17]

10. See references [6, 13]

QUALITATIVE SYNTHESIS OF NONLINEAR NETWORKS

RUEY-WEN LIU and LEO V. AUTH, JR.
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana

Abstract

Qualitative synthesis is defined in terms of qualitative properties of the solutions of nonlinear differential equations. Liapunov's direct method is used to find sufficient conditions on the element voltage-current relations and network topology so that the qualitative properties of the solutions resemble those of strongly stable linear networks. Two theorems concerning complete stability and ultimate boundedness are obtained.

Introduction

The realization step of linear network synthesis normally begins with one or more of the various network functions specified in terms of a complex frequency variable and ends with the determination of the network element values subject to constraints on the network topology and kinds of elements to be utilized. The network functions have presumably been obtained by approximation of desired performance criteria. In the linear domain, the network function completely determines the output response of the network, when it is in an initially relaxed state, to an arbitrary input excitation. Superposition then extends this knowledge to any number of inputs either at the same port or combination of other ports. Furthermore, the important questions of stability and boundedness of the output for a bounded input are determined by the locations of the poles of the pertinent network functions. The existence of the network functions and the consequent relative ease of the solution of the linear network synthesis problem rest primarily upon the properties and form of the solutions of the linear differential equations which describe these networks. In the nonlinear case, however, the differential equations governing the response of the network do not have such desirable properties. Indeed, general methods for obtaining explicit solutions are unknown. Thus, the complete characterization of a general nonlinear network by quantities analogous to the familiar network functions appears highly improbable. Moreover, the synthesis result of explicit functional characteristics of nonlinear elements to give input-output relations prescribed in a quantitative way seems too ambitious at this time.

Qualitative Behavior of Solutions

We begin by considering the qualitative properties of the solution of a set of differential equations expressed in normal form

$$\dot{x}_i = f_i(x_1, x_2, \ldots, x_n; t), \quad i = 1, 2, \ldots, n \tag{1}$$

where the dot denotes the time derivative. Eq. (1) can be written in the vector form

$$\dot{\underset{\sim}{x}} = \underset{\sim}{f}(\underset{\sim}{x}, t) \tag{2}$$

where $\underset{\sim}{x}$, $\underset{\sim}{f}$ are n-vectors with components x_i and f_i respectively. Let $\underset{\sim}{f}$ satisfy the conditions[1] so that the solution exists and is unique, i.e., for any given initial condition $(\underset{\sim}{x}_o, t_o)$, there exists a unique solution $\underset{\sim}{x} = \underset{\sim}{\Phi}(t; \underset{\sim}{x}_o, t_o)$ such that the initial condition $\underset{\sim}{x}_o = \underset{\sim}{\Phi}(t_o; \underset{\sim}{x}_o, t_o)$ is satisfied.

In general, the solution can be qualitatively classified as follows:

I. Unbounded:

(a) $\|\Phi\| \rightarrow \infty$ as $t \rightarrow T$, where $t_o < T < \infty$: The solution is unbounded after a finite time. This solution is said to have finite escape time. For example, the solution of Van der Pol's equation

$$\ddot{x} - \epsilon(1 - x^2)\dot{x} + x = f(t) \text{ with } \epsilon < 0 \text{ and}$$

bounded f(t) may have finite escape time.[2]

(b) $\|\Phi\| \rightarrow \infty$ as $t \rightarrow \infty$. Unbounded solutions of linear systems are of this type.

II. Bounded:

(c) $\underset{\sim}{\Phi} = \underset{\sim}{C}$ for all t, where $\underset{\sim}{C}$ is a constant vector: if the solution is initially at $\underset{\sim}{C}$ it stays at $\underset{\sim}{C}$. Hence $\underset{\sim}{x} = \underset{\sim}{C}$ is an equilibrium position.

(d) $\underset{\sim}{\Phi}$ is periodic: $\underset{\sim}{\Phi}$ is a bounded periodic function. It is noted that an unbounded periodic function cannot be a solution. In fact, it would have finite escape time.

(e) $\underset{\sim}{\Phi}$ approaches an equilibrium position as $t \rightarrow \infty$.

(f) $\underset{\sim}{\Phi}$ approaches a bounded periodic function as $t \rightarrow \infty$.

(g) $\underset{\sim}{\Phi}$ is bounded but does not belong to (c), (d) (e) and (f): this includes almost periodic solutions and solutions which approach them.

It is remarked that the qualitative behavior of linear systems is generally solely determined by the differential equation itself. This is not true for nonlinear differential equations. For example, the solution of $\dot{x} = -x^2$ is bounded if initially $x_o \geqq 0$ and unbounded if $x_o < 0$.

Definition of Qualitative Synthesis

Given a set of input-output relationships, the problem of synthesis is to find the system to give the prescribed conditions. The problem of qualitative synthesis of nonlinear networks is to find the topology of the network and the nonlinear characteristics of the individual elements when the input-output relations are prescribed in a qualitative way.

This paper is concerned with only the following two cases out of the many possible combinations from the list of qualitative properties as given above.

Input	Output
1) Constant zero input	All solutions approach $\underset{\sim}{x} = \underset{\sim}{0}$
2) Bounded input	All solutions are bounded

For linear networks these two cases are equivalent and such networks are classified as strongly stable. It will be shown that some nonlinear networks will satisfy one but not the other.

Primary Definitions and Theorems

The problem of synthesis is carried out by Liapunov's Direct Method. Liapunov [3] introduced the method to study stability problems in 1893, Yoshizawa [4,5] extended the method to study the problem of boundedness and Skowronski [6,7] applied it to mainly mechanical systems. Two excellent references [1,2] concerning the direct method are available.

Because of the fact that the qualitative properties of the solutions depend on the initial conditions, many different types of stability and boundedness have been defined. A few of those commonly used are introduced here. These definitions and others can be found in [1,2].

Let

$$\|\underset{\sim}{x}\| = (\Sigma x_1^2)^{\frac{1}{2}}$$

denote the distance from $\underset{\sim}{x}$ to the origin $\underset{\sim}{0}$. Then

$$\|\underset{\sim}{x}\| < r$$

is the spherical region centered at $\underset{\sim}{0}$ with radius r, which is denoted by S(r).

Consider again equation (2) with $\underset{\sim}{f}(\underset{\sim}{0},t) = 0$.[3] Hence $\underset{\sim}{x} = \underset{\sim}{0}$ is an equilibrium position, i.e., $\underset{\sim}{x} = \underset{\sim}{0}$ for all t is a solution.

Definition 1: The equilibrium position $\underset{\sim}{x} = \underset{\sim}{0}$ of (2) is stable if for every $\epsilon > 0$, there exists $\delta > 0$, such that if

$$\|\underset{\sim}{x}_o\| < \delta$$

then

$$\|\underset{\sim}{\Phi}(t,\underset{\sim}{x}_o,t_o)\| < \epsilon \qquad \text{for all } t \geqq t_o$$

The latter two inequalities mean that if the initial position is inside the spherical region $S(\delta)$, then the consequent solution remains in $S(\epsilon)$.

Definition 2: The $\underset{\sim}{x} = \underset{\sim}{0}$ is asymptotically stable if it is stable and there exists $R_o > 0$ such that

$$\lim \underset{\sim}{\Phi}(t,\underset{\sim}{x}_o,t_o) \longrightarrow \underset{\sim}{0}$$

if $\underset{\sim}{x}_o$ is initially inside the spherical region $S(R_o)$. This spherical region is called the region of asymptotic stability.

For example, an LC connected in parallel is stable but not asymptotically stable, since both the current and the voltage will stay near $\underset{\sim}{0}$ if they are near $\underset{\sim}{0}$ initially, but will not approach $\underset{\sim}{0}$.

It is noted that, with the assumption that infinitesmal disturbances always exist, an unstable equilibrium position in general cannot be obtained experimentally.[4]

These two definitions concern the stability of the equilibrium position. They shall not be confused with the stability of a network which concerns all solutions. The following definition by La Salle[5] concerns the latter problem.

Definition 3: If $\underset{\sim}{x} = \underset{\sim}{0}$ is asymptotically stable and the whole space is the region of stability, then we have complete stability.

Hence complete stability implies that all solutions approach $\underset{\sim}{0}$. A linear system with no excitation and the characteristic roots in the left half plane is completely stable.

Finally, a definition of boundedness is introduced here.

Definition 4: The solutions of (2) are ultimately bounded if there is a bound b such that for every solution one has a $\tau > 0$ and

$$\left\|\underset{\sim}{\Phi}(t,\underset{\sim}{x}_o,t_o)\right\| < b \text{ for } t > t_o + \tau$$

This implies that every solution is bounded by the spherical region S(b) after τ second. For linear networks, this corresponds to the steady-state being bounded by a definite value b. A linear system with all characteristic roots in the left half plane and with a bounded excitation is ultimately bounded.

The following theorems give the sufficient conditions of complete stability and ultimate boundedness. The first theorem concerns the autonomous system.

$$\dot{\underset{\sim}{x}} = \underset{\sim}{f}(\underset{\sim}{x}); \quad \underset{\sim}{f}(\underset{\sim}{0}) = \underset{\sim}{0} \tag{3}$$

i.e., $\underset{\sim}{f}$ is not a function of t in (2). An autonomous system implies nonexistence of independent sources.

Theorem I: If there exists a function $V(\underset{\sim}{x})$ such that

(I - 1) $V(\underset{\sim}{x})$ in positive definite

(I - 2) $V(\underset{\sim}{x}) \longrightarrow \infty$ as $\left\|\underset{\sim}{x}\right\| \longrightarrow \infty$

(I - 3) $V = \underset{\sim}{f} \cdot V\,V < 0$ for all $\underset{\sim}{x} \neq 0$

then System (3) is completely stable.

Theorem II: If there exists a function V(x) and $\epsilon > 0$, such that

(II-1) $V(\underset{\sim}{x}) \longrightarrow \infty$ as $\left\|\underset{\sim}{x}\right\| \longrightarrow \infty$

(II-2) $\dot{V}(\underset{\sim}{x}) \leq -\epsilon$ for all $t \geq 0$ and for all $\underset{\sim}{x}$ outside some closed and bounded set M,

then System (2) is ultimately bounded.

The proofs of these theorems are found in [1,2]. It is remarked that if 2) of Theorem I is not satisfied, then it is sufficient that the equilibrium position of System (3) is asymptotically stable. The function $V(\underset{\sim}{x})$ is called the Liapunov function. The advantage of this method is that information of qualitative properties can be obtained without solving the differential equation. The difficulty lies at finding a $V(\underset{\sim}{x})$ satisfying the respective conditions.

Existence of the Normal Form for a Subclass of Nonlinear Networks

The behavior of any electrical network is completely determined when the voltage across and current through every branch of the network are known for all time. For the linear case, these 2b variables, i.e., b branch voltages $\underset{\sim}{v}$ and b branch currents $\underset{\sim}{i}$, are not dynamically independent and at least one subset exists in terms of which all remaining variables are uniquely determined. Such a subset is called a set of state variables and Bryant [8] has recently given the explicit form for the solution of a linear network in terms of a particular set of its state variables. The state variables chosen by Bryant are the currents through inductors $\underset{\sim}{i}_\gamma$ forming a set of chords for a particular tree of the network and the voltages across capacitors $\underset{\sim}{v}_\delta$ present in the same tree. Inductors and capacitors so chosen represent a set of independent points of energy storage for the network. Elimination of non-dynamically independent variables from the general 2b equations is accomplished by using algebraic constraints arising from Kirchhoff's Laws Ohm's Law, capacitor-only tie sets, and inductor-only cut sets. The final result achieved by Bryant in normal form is

$$\frac{d}{dt}\begin{bmatrix} \underset{\sim}{i}_\gamma \\ \\ \underset{\sim}{v}_\delta \end{bmatrix} = -\underset{\sim}{M}^{-1}\,\underset{\sim}{N}\begin{bmatrix} \underset{\sim}{i}_\gamma \\ \\ \underset{\sim}{v}_\delta \end{bmatrix} + \underset{\sim}{M}^{-1}\,\underset{\sim}{y} \qquad (4)$$

where the matrices $\underset{\sim}{M}$, $\underset{\sim}{N}$, and $\underset{\sim}{y}$ are determined by the network topology and element values with $\underset{\sim}{y}$ additionally involving the independent sources i.e., $\underset{\sim}{y}$ is the forcing matrix. Since all theorems concerning stability and boundedness for solutions of nonlinear differential equations premise the existence of the equations in normal form, it is necessary to establish the existence of this form for the nonlinear networks under consideration although it should be recalled that the solution of such equations cannot be accomplished in general and indeed need not be known for the determination of qualitative properties.

In this paper, we shall consider nonlinear networks composed of two terminal elements whose characteristics are time-invariant and whose branch voltage and branch current for an individual element are related in one of the following ways:

$$(a)\quad v_R = f_R(i_R) = R(i_R)i_R \quad \text{or} \quad i_R = g_R(v_R) = G(v_R)v_R$$

$$(b)\quad i_C = \frac{dq_C}{dt} = \frac{dq_C}{dv_C}\,\frac{dv_C}{dt} = C(v_C)\,\frac{dv_C}{dt}$$

$$(c)\quad v_L = \frac{d\psi_L}{dt} = \frac{d\psi_{LL}}{di_L}\,\frac{di_L}{dt} + \sum_{j\neq L} \frac{d\psi_{Lj}}{di_j}\,\frac{di_j}{dt} \qquad (5)$$

$$= L_{LL}\,(i_L)\,\frac{di_L}{dt} + \sum_{j\neq L} M_{Lj}(i_j)\,\frac{di_j}{dt}$$

and where we further assume that the functions f, q, and ψ are zero when their arguments are zero[6] and the partial derivatives exist and are continuous and single-valued with respect to the variable involved. Thus we do not consider hysteresis effects nor allow the presence of controlled sources where in the latter one might have $i = g(e_1, e_2, \ldots)$.

In the Appendix it is shown that, with certain other conditions on the network elements, the normal form exists for nonlinear networks in terms of the same set of $\underset{\sim}{i}_\gamma$ and $\underset{\sim}{v}_\delta$ as used by Bryant. These conditions are

1) $f_R(i_R)$ is a monotonic increasing function
2) $\underset{\sim}{C}$ and $\underset{\sim}{L}$ are definite matrices throughout the state space.

The explicit equation in final normal form, analogous to (4) for the linear case, cannot be shown.

Nonlinear Network Theorems

We assume that independent voltage and current sources present in the network occur in the general manner adopted by Bryant.[7] Under these conditions KCL is given by

$$\underset{\sim}{A}\,(\underset{\sim}{i} - \underset{\sim}{j}) = \underset{\sim}{0}$$

and KVL by

$$\underset{\sim}{B}\,(\underset{\sim}{v} - \underset{\sim}{e}) = \underset{\sim}{0}$$

where $\underset{\sim}{j}$ and $\underset{\sim}{e}$ are the column matrices representing the respective sources which areonly dependent on time (t). A well known result[8] of the branch voltages and currents satisfying KCL and KVL is that

$$\underset{\sim}{i}'\underset{\sim}{v} - \underset{\sim}{j}'\underset{\sim}{v} - \underset{\sim}{i}'\underset{\sim}{e} + \underset{\sim}{j}'\underset{\sim}{e} = 0$$

Using (5) and the chord-branch classification of Bryant,[9] (6) may be reduced to

$$\begin{bmatrix}\dot{\underset{\sim}{v}}'_{\alpha} & \dot{\underset{\sim}{v}}'_{\delta}\end{bmatrix}\begin{bmatrix}\underset{\sim}{C}_{\alpha\alpha} & \underset{\sim}{0}\\ \underset{\sim}{0} & \underset{\sim}{C}_{\delta\delta}\end{bmatrix}\begin{bmatrix}\underset{\sim}{v}_{\alpha}\\ \underset{\sim}{v}_{\delta}\end{bmatrix} + \begin{bmatrix}\underset{\sim}{i}'_{\beta} & \underset{\sim}{i}'_{\epsilon}\end{bmatrix}\begin{bmatrix}\underset{\sim}{R}_{\beta\beta} & \underset{\sim}{0}\\ \underset{\sim}{0} & \underset{\sim}{R}_{\epsilon\epsilon}\end{bmatrix}\begin{bmatrix}\underset{\sim}{i}_{\beta}\\ \underset{\sim}{i}_{\epsilon}\end{bmatrix}$$

$$+ \begin{bmatrix}\underset{\sim}{i}'_{\gamma} & \underset{\sim}{i}'_{\zeta}\end{bmatrix}\begin{bmatrix}\underset{\sim}{L}_{\gamma\gamma} & \underset{\sim}{L}_{\gamma\zeta}\\ \underset{\sim}{L}_{\zeta\gamma} & \underset{\sim}{L}_{\zeta\zeta}\end{bmatrix}\begin{bmatrix}\dot{\underset{\sim}{i}}_{\gamma}\\ \dot{\underset{\sim}{i}}_{\zeta}\end{bmatrix} = \underset{\sim}{j}'\underset{\sim}{v} + \underset{\sim}{i}'\underset{\sim}{e} - \underset{\sim}{j}'\underset{\sim}{e} \qquad (7)$$

Let us now apply Liapunov's direct method to the network problem. In general there is no definite way for choosing the function $V(\underset{\sim}{x})$, however, a natural function to choose will be the total energy stored in the reactive elements,

$$V = \int_{0}^{\underset{\sim}{i}} \begin{bmatrix}\underset{\sim}{i}'_{\gamma} & \underset{\sim}{i}'_{\zeta}\end{bmatrix}\begin{bmatrix}\underset{\sim}{L}_{\gamma\gamma} & \underset{\sim}{L}_{\gamma\zeta}\\ \underset{\sim}{L}_{\zeta\gamma} & \underset{\sim}{L}_{\zeta\zeta}\end{bmatrix}\begin{bmatrix}d\underset{\sim}{i}_{\gamma}\\ d\underset{\sim}{i}_{\zeta}\end{bmatrix} \qquad (8)$$

$$+ \int_{0}^{\underset{\sim}{v}} \begin{bmatrix}d\underset{\sim}{v}'_{\alpha} & d\underset{\sim}{v}'_{\delta}\end{bmatrix}\begin{bmatrix}\underset{\sim}{C}_{\alpha\alpha} & \underset{\sim}{0}\\ \underset{\sim}{0} & \underset{\sim}{C}_{\delta\delta}\end{bmatrix}\begin{bmatrix}\underset{\sim}{v}_{\alpha}\\ \underset{\sim}{v}_{\delta}\end{bmatrix}$$

thereby giving

$$\dot{V} = -\begin{bmatrix}\underset{\sim}{i}'_{\beta} & \underset{\sim}{i}'_{\epsilon}\end{bmatrix}\begin{bmatrix}\underset{\sim}{R}_{\beta\beta} & \underset{\sim}{0}\\ \underset{\sim}{0} & \underset{\sim}{R}_{\epsilon\epsilon}\end{bmatrix}\begin{bmatrix}\underset{\sim}{i}_{\beta}\\ \underset{\sim}{i}_{\epsilon}\end{bmatrix} + \underset{\sim}{j}'\underset{\sim}{v} + \underset{\sim}{i}'\underset{\sim}{e} - \underset{\sim}{j}'\underset{\sim}{e} \qquad (9)$$

Thus $\dot{V}$ is the negative of the total net power dissipation of the network.

The problem is now to put restrictions on the element functional relations and the topology of the network so that the conditions in the respective theorems are satisfied. It appears that the test for definiteness of the form of (8) has not been established when $\underset{\sim}{L}$ is not a diagonal matrix and its elements are not constants. Therefore, for simplicity of the statements of the following theorems, we further assume no mutual inductive coupling. In this case, the positive definiteness of $\underset{\sim}{L}$, which is now diagonal, and $\underset{\sim}{C}$ ensures the positive definiteness of (8). Sets of sufficient conditions to give complete stability and ultimate boundedness are obtained and are given in the next two theorems.

Theorem III:

The network with no independent sources is completely stable if the following conditions are satisfied:

(III - 1) The diagonal matrices $\underset{\sim}{L}$ and $\underset{\sim}{C}$ are positive definite.

(III - 2) $f_R(i_R)$ is a monotonic and increasing function

(III - 3) $$\int_{\underset{\sim}{0}}^{\underset{\sim}{i}_\gamma} \underset{\sim}{i}'_\gamma \, \underset{\sim}{L}_{\gamma\gamma}(\underset{\sim}{i}_\gamma) \, d\underset{\sim}{i}_\gamma \longrightarrow \infty \quad \text{as} \quad \|\underset{\sim}{i}_\gamma\| \longrightarrow \infty$$

$$\int_{\underset{\sim}{0}}^{\underset{\sim}{v}_\delta} \underset{\sim}{v}'_\delta \, \underset{\sim}{C}_{\delta\delta}(\underset{\sim}{v}_\delta) \, d\underset{\sim}{v}_\delta \longrightarrow \infty \quad \text{as} \quad \|\underset{\sim}{v}_\delta\| \longrightarrow \infty$$

(III - 4) The state variables $(\underset{\sim}{v}_\delta, \underset{\sim}{i}_\gamma)$ are zero if and only if all currents in resistive elements are zero.

Proof: As shown in the appendix, (III - 1, 2) established that $(\underset{\sim}{v}_\delta, \underset{\sim}{i}_\gamma)$ are the state variables $(\underset{\sim}{x})$. Further, $V(\underset{\sim}{x})$ is positive definite because of (III -1). Condition (III - 3) insures the satisfaction of (I -2). From (9) and (III - 2), $\dot{V}$ is non-positive and equal to zero only if all the resistive currents are zero by (III - 4). Hence (I - 3) is fulfilled and the theorem is proved.

It is noted that (III - 3) implies that the energy stored in the reactive elements cannot be finite when $\underset{\sim}{i}_\gamma$ or $\underset{\sim}{v}_\delta$ approaches infinity, a reasonable requirement that is satisfied by all practical reactive elements. Condition (III - 4) depends on the topology of the network. For example, it will be satisfied, if there are resistive elements in series with the chord inductive elements and in parallel with the branch capacitive elements. Another sufficient condition is that the rank of the matrix[10]

$$\begin{bmatrix} -\underset{\sim}{F}_{\beta\delta} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{F}'_{\gamma\epsilon} \end{bmatrix}$$

is no less than $\delta + \gamma$.

In addition one may put more restrictions on the topology of the network so that the restrictions on the elements can be relaxed. For example, if the state variables can be obtained without the monotonic characteristic for the resistive elements, then (III - 2) can be relaxed to $i_R f_R(i_R) > 0$ for $i_R \neq 0$; hence one may admit regions of negative slopes.

It is noted that a bounded input to a completely stable network does not always produce a bounded output. For example, consider the equation

$$\ddot{q} + f(\dot{q}) + q = 10 \sin t$$

where $f(\dot{q})$ is shown in Fig. 1. This equation may be considered as linear L and C in series with a limiter. It can be shown[11] that the solution is not bounded although the network without the input is completely stable.

The following theorem gives the sufficient conditions for ultimate boundedness.

Theorem IV: With any bounded input, the output of the network is ultimately bounded if the following conditions are satisfied:

(IV - 1) (III-1,2,3) are satisfied

(IV - 2) $f_R(i_R) \longrightarrow \infty$ as $i_R \longrightarrow \infty$

(IV - 3) each e(t) is in series with a resistive element and each j(t) in parallel with a resistive element.

(IV - 4) $\left\| (\underset{\sim}{v}_\delta, \underset{\sim}{i}_\gamma) \right\| \longrightarrow \infty$ implies that $\left\| (\underset{\sim}{i}_\beta, \underset{\sim}{i}_\epsilon) \right\| \longrightarrow \infty$

Proof: For the same reason as in the proof of Theorem III, (II - 1) is satisfied. For large current, the power dissipation in the resistive element, in view of (IV -2,3), is larger than the power generated by the voltage generator in series with it and the current generator in parallel with it. In fact, the difference can be arbitrarily large for large current. Since - $\dot{V}$ is the total net power dissipation in the network, $\dot{V}$ is negative and does not approach zero for $\left\| (\underset{\sim}{i}_\beta, \underset{\sim}{i}_\epsilon) \right\|$ large enough; therefore, because of (IV - 4) $\dot{V}$ is negative and does not approach zero for $\left\| (\underset{\sim}{v}_\delta, \underset{\sim}{i}_\gamma) \right\|$ large enough. The theorem is proved.

It is seen that the above example shows the necessity of (IV - 2).

It is remarked that (IV - 4) again depends on the topology of the network, and it will be satisfied, for example, if there are resistive elements in series with the chord inductive elements and in parallel with the branch capacitive elements. Finally, it is emphasized that the conditions given in Theorems III and IV are sufficient conditions only. For a different assumed function V, a different set of sufficient conditions may be obtained.

Conclusion

The results obtained serve to indicate the applicability of Liapunov's direct method to the problem of qualitative synthesis of nonlinear networks. In the derivation of the network theorems it is seen that restrictions on the network topology and voltage-current relations of the elements arise naturally in meeting the conditions of the stability and boundedness theorems. This result is not unlike the linear cases where it is common to restrict the topology and kinds of allowable elements at the outset of the synthesis procedure

Although the direct method has been applied to only two special cases, it can be applied to almost all combinations of the properties given in the Introduction. Different expressions of the Liapunov function can also be tried to study the synthesis problems. In particular, modifications of a useful Liapunov function may lead to semi-quantitative results, when, for example, a numerical value is provided for the bound of the response of the desired network.

References

[1] J. La Salle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, 1961.

[2] W. Hahn, *Theory and Application of Liapunov's Direct Method*, (English translation), Prentice-Hall, 1963.

[3] M. A. Liapunov, "Probleme general de la stabilite du mouvement" *Ann. Fac. Sci. Toulouse,* Vol. 9, pp. 203-474, 1907. (French translation of the original paper published in 1893 in *Comm. Soc. Math. Kharkow*; reprinted as Vol. 17 in *Ann. Math. Studies*, Princeton 1949).

[4] T. Yoshizawa, "Note on the boundedness of solutions of a system of differential equations", *Mem. Coll. Sci. Kyoto*, Ser. A, Vol. 28, pp. 203-298, 1954.

[5] T. Yoshizawa, "Liapunov's functions and boundedness of solutions", *Funkcialaj Ekvacioj*, Vol. 2, pp. 95-142, 1959.

[6] J. M. Skowronski and S. Ziemba, "Boundedness of motion and existence and stability of the limit regime in strongly nonlinear non-autonomous vibrating discrete systems", *Proc. of Intern. Congress of I F A C*, Moscow, 1960.

[7] J. M. Skowronski, "Character of motion of the strongly nonlinear mechanical systems", *Zagadnienia drgan nieliniowych*, Warsaw, Vol 4, pp. 4-52, 1962.

[8] P. R. Bryant, "The explicit form of Bashkow's A matrix", *IRE Trans. on Circuit Theory*, Vol CT-9, pp. 303-306, September, 1962.

[9] P. R. Bryant, "The order of complexity of electrical networks", *Proc. IEE*, Monograph 335E, Vol 106C, pp. 174-188, June, 1959.

[10] S. Seshu and M. B. Reed, *Linear Graphs and Electrical Networks*, Addison-Wesley Publishing Company, Inc., Reading, Mass., 1961.

Appendix

We adopt the notation used by Bryant in [8] and [9]. It is evident that Bryant's $(7)^{12}$, which is a single matrix equation combining KCL KVL and the element voltage-current relations, remains valid for networks containing the nonlinear elements defined by (5) after replacing $\underset{\sim}{R}$ by $\underset{\sim}{R}(\underset{\sim}{i})$, $\underset{\sim}{C}$ by $\underset{\sim}{C}(\underset{\sim}{v})$ and $\underset{\sim}{L}$ by $\underset{\sim}{L}(\underset{\sim}{i})$. Assuming that the algebraic manipulations used in the elimination process would also be applicable to the nonlinear case, his final result (4) would no longer be in normal form since the elements of $\underset{\sim}{M}$, $\underset{\sim}{N}$, and $\underset{\sim}{y}$ would be functions of the non-state-variables. Therefore, it is necessary to critically examine the conditions for the existence of unique solutions to the algebraic equations used in the elimination process. This process proceeds in three steps; elimination using algebraic constraints from

1. Kirchhoff's and Ohm's Law
2. Capacitor-only tie sets
3. Inductor-only cut sets.

Case 1:

Bryant's (77)[13] becomes for the nonlinear case

$$\begin{aligned} \underset{\sim}{f}_\beta(\underset{\sim}{i}_\beta) + \underset{\sim}{F}_{\beta\epsilon}\underset{\sim}{v}_\epsilon &= -\underset{\sim}{F}_{\beta\delta}\underset{\sim}{v}_\delta + \underset{\sim}{E}_\beta \\ -\underset{\sim}{F}'_{\beta\epsilon}\underset{\sim}{i}_\beta + \underset{\sim}{g}_\epsilon(\underset{\sim}{v}_\epsilon) &= \underset{\sim}{F}'_{\gamma\epsilon}\underset{\sim}{i}_\gamma + \underset{\sim}{T}_\epsilon \end{aligned} \tag{10}$$

Let

$$\underset{\sim}{y}_\beta = -\underset{\sim}{F}_{\beta\delta}\underset{\sim}{v}_\delta + \underset{\sim}{E}_\beta$$

$$\underset{\sim}{y}_\epsilon = \underset{\sim}{F}'_{\gamma\epsilon}\underset{\sim}{i}_\gamma + \underset{\sim}{T}_\epsilon$$

We need to show that, for a given $(\underset{\sim}{v}_\delta, \underset{\sim}{i}_\gamma)$, a set $(\underset{\sim}{i}_\beta, \underset{\sim}{v}_\epsilon)$ is uniquely determined. It is seen that for any $(\underset{\sim}{v}_\delta, \underset{\sim}{i}_\gamma)$, the pair $(\underset{\sim}{y}_\beta, \underset{\sim}{y}_\epsilon)$ is uniquely determined ($\underset{\sim}{E}_\beta$ and $\underset{\sim}{T}_\epsilon$ represent independent sources), hence we only need to show that $(\underset{\sim}{i}_\beta, \underset{\sim}{v}_\epsilon)$ is uniquely determined for a given set $(\underset{\sim}{y}_\beta, \underset{\sim}{y}_\epsilon)$. A sufficient condition for this to be true is that $\underset{\sim}{f}_\beta$ and $\underset{\sim}{g}_\epsilon$ are monotinic functions. The proof is by contradiction:

If there are two sets $(\underset{\sim}{i}_{\beta 1}, \underset{\sim}{y}_{\epsilon 1})$ and $(\underset{\sim}{i}_{\beta 2}, \underset{\sim}{y}_{\epsilon 2})$ which map to the same $(\underset{\sim}{y}_\beta, \underset{\sim}{y}_\epsilon)$, then we have

$$\underset{\sim}{k}_{\beta\beta}\Delta\underset{\sim}{i}_\beta + \underset{\sim}{F}_{\beta\epsilon}\Delta\underset{\sim}{v}_\epsilon = \underset{\sim}{0}$$

$$-\underset{\sim}{F}'_{\beta\epsilon}\Delta\underset{\sim}{i}_\beta + \underset{\sim}{h}_{\epsilon\epsilon}\Delta\underset{\sim}{v}_\epsilon = \underset{\sim}{0}$$

where $\Delta\underset{\sim}{i}_\beta = \underset{\sim}{i}_{\beta 1} - \underset{\sim}{i}_{\beta 2}$ and $\Delta\underset{\sim}{v}_\epsilon = \underset{\sim}{v}_{\epsilon 1} - \underset{\sim}{v}_{\epsilon 2}$ and $\underset{\sim}{k}_{\beta\beta}$ and $\underset{\sim}{h}_{\epsilon\epsilon}$ have positive elements because of the monotonicity property of $\underset{\sim}{f}_\beta$ and $\underset{\sim}{g}_\epsilon$. Since $\underset{\sim}{k}_{\beta\beta}^{-1}$ and $\underset{\sim}{h}_{\epsilon\epsilon}^{-1}$ exist (they are diagonal matrices with nonzero diagonal entries) we have

$$\Delta\underset{\sim}{i}_\beta + \underset{\sim}{k}_{\beta\beta}^{-1}\underset{\sim}{F}_{\beta\epsilon}\Delta\underset{\sim}{v}_\epsilon = \underset{\sim}{0}$$

$$-\underset{\sim}{h}_{\epsilon\epsilon}^{-1}\underset{\sim}{F}'_{\beta\epsilon}\Delta\underset{\sim}{i}_\beta + \Delta\underset{\sim}{v}_\epsilon = \underset{\sim}{0}$$

Since

$$\det\begin{bmatrix} \underset{\sim}{1}_{\beta\beta} & \underset{\sim}{k}_{\beta\beta}^{-1}\underset{\sim}{F}_{\beta\epsilon} \\ -\underset{\sim}{h}_{\epsilon\epsilon}^{-1}\underset{\sim}{F}'_{\beta\epsilon} & \underset{\sim}{1}_{\epsilon\epsilon} \end{bmatrix} = \det\begin{bmatrix} \underset{\sim}{k}_{\beta\beta}^{-1}\begin{bmatrix}\underset{\sim}{F}_{\beta\epsilon} & \underset{\sim}{1}_{\beta\beta}\end{bmatrix} \begin{bmatrix} \underset{\sim}{h}_{\epsilon\epsilon}^{-1} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{k}_{\beta\beta} \end{bmatrix}\end{bmatrix}\begin{bmatrix} \underset{\sim}{F}'_{\beta\epsilon} \\ \underset{\sim}{1}_{\beta\beta} \end{bmatrix}$$

then the determinant is nonzero[15] since

$$\begin{bmatrix} \underset{\sim}{h}_{\epsilon\epsilon}^{-1} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{k}_{\beta\beta} \end{bmatrix}$$

is positive definite, again by the monotonicity of $\underset{\sim}{f}_{\beta}$ and $\underset{\sim}{g}_{\epsilon}$, hence $\Delta\underset{\sim}{i}_{\beta} = \underset{\sim}{0}$ and $\Delta\underset{\sim}{v}_{\epsilon} = \underset{\sim}{0}$.

It is noted that the relation between $(\underset{\sim}{v}_{\delta}, \underset{\sim}{i}_{\gamma})$ and $(\underset{\sim}{i}_{\beta}, \underset{\sim}{v}_{\epsilon})$ may not cover the entire range of values. For example, if a limiter is present, the corresponding variable will be bounded. However, this relation will exist for all values that are possible. A system completely stable for all possible values is sometimes called asymptotically stable in the large.

Case 2:

The nonlinear equations analogous to Bryant's (89)[14] are

$$\begin{bmatrix} \underset{\sim}{F}_{\alpha\delta} & \underset{\sim}{1}_{\alpha\alpha} \end{bmatrix} \begin{bmatrix} \underset{\sim}{C}_{\delta\delta}^{-1}(\underset{\sim}{v}_{\delta}) & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{C}_{\alpha\alpha}^{-1}(\underset{\sim}{v}_{\alpha}) \end{bmatrix} \begin{bmatrix} \underset{\sim}{F}'_{\alpha\delta} \\ \underset{\sim}{1}_{\alpha\alpha} \end{bmatrix} \underset{\sim}{i}_{\alpha} =$$

$$\dot{\underset{\sim}{E}}_{\alpha} - \underset{\sim}{F}_{\alpha\delta}\underset{\sim}{C}_{\delta\delta}^{-1}(\underset{\sim}{v}_{\delta}) \begin{bmatrix} \underset{\sim}{F}'_{\beta\delta}\underset{\sim}{i}_{\beta} + \underset{\sim}{F}_{\gamma\delta}\underset{\sim}{i}_{\gamma} + \underset{\sim}{T}_{\delta} \end{bmatrix} \tag{11}$$

and

$$\underset{\sim}{v}_{\alpha} = -\underset{\sim}{F}_{\alpha\delta}\underset{\sim}{v}_{\delta} + \underset{\sim}{E}_{\alpha}$$

Equation (11) is a linear equation in i_{α} and a unique solution exists if

$$\det \begin{bmatrix} \begin{bmatrix} \underset{\sim}{F}_{\alpha\delta} & \underset{\sim}{1}_{\alpha\alpha} \end{bmatrix} \begin{bmatrix} \underset{\sim}{C}_{\delta\delta}^{-1}(\underset{\sim}{v}_{\delta}) & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{C}_{\alpha\alpha}^{-1}(\underset{\sim}{v}_{\alpha}) \end{bmatrix} \begin{bmatrix} \underset{\sim}{F}'_{\alpha\delta} \\ \underset{\sim}{1}_{\alpha\alpha} \end{bmatrix} \end{bmatrix}$$

is nonsingular. A sufficient condition is then that

$$\begin{bmatrix} \underset{\sim}{C}_{\delta\delta}^{-1}(\underset{\sim}{v}_{\delta}) & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{C}_{\alpha\alpha}^{-1}(\underset{\sim}{v}_{\alpha}) \end{bmatrix}$$

is a definite matrix throughout the state space.

Case 3:

If we define

$$\underset{\sim}{L}^{-1} = \begin{bmatrix} \underset{\sim}{L}_{\gamma\gamma} & (\underset{\sim}{i}_{\gamma}) & \underset{\sim}{L}_{\gamma\zeta} & (\underset{\sim}{i}_{\zeta}) \\ \underset{\sim}{L}_{\zeta\gamma} & (\underset{\sim}{i}_{\gamma}) & \underset{\sim}{L}_{\zeta\zeta} & (\underset{\sim}{i}_{\zeta}) \end{bmatrix}^{-1} = \begin{bmatrix} \underset{\sim}{T}_{\gamma\gamma}(\underset{\sim}{i}_{\gamma}, \underset{\sim}{i}_{\zeta}) & \underset{\sim}{T}_{\gamma\zeta}(\underset{\sim}{i}_{\gamma}, \underset{\sim}{i}_{\zeta}) \\ \underset{\sim}{T}_{\zeta\gamma}(\underset{\sim}{i}_{\gamma}, \underset{\sim}{i}_{\zeta}) & \underset{\sim}{T}_{\zeta\zeta}(\underset{\sim}{i}_{\gamma}, \underset{\sim}{i}_{\zeta}) \end{bmatrix} = \underset{\sim}{T}$$

Then the equations for v_{ζ} in the nonlinear case becomes

$$\begin{bmatrix} \underset{\sim}{F}'_{\gamma\delta} & \underset{\sim}{i}_{\zeta\zeta} \end{bmatrix} \begin{bmatrix} \underset{\sim}{T} \end{bmatrix} \begin{bmatrix} \underset{\sim}{F}_{\gamma\zeta} \\ \underset{\sim}{1}_{\zeta\zeta} \end{bmatrix} \begin{bmatrix} \underset{\sim}{v}_{\zeta} \end{bmatrix} = \dot{\underset{\sim}{T}}_{\zeta} + \begin{bmatrix} \underset{\sim}{T}_{\zeta\gamma} & -\underset{\sim}{F}'_{\gamma\zeta}\underset{\sim}{T}_{\gamma\gamma} \end{bmatrix} \begin{bmatrix} \underset{\sim}{F}_{\gamma\delta}\underset{\sim}{v}_{\delta} + \underset{\sim}{F}_{\gamma\epsilon}\underset{\sim}{v}_{\epsilon} + \underset{\sim}{E}_{\gamma} \end{bmatrix} \tag{13}$$

and

$$\underset{\sim}{i}_{\zeta} = \underset{\sim}{F}'_{\gamma\zeta}\underset{\sim}{i}_{\gamma} + \underset{\sim}{T}_{\zeta} \tag{14}$$

Equation (13) is a linear equation in $\underset{\sim}{v}_{\zeta}$ and the existence of a unique solution is again ensured by the sufficient condition that $\underset{\sim}{T}$, and hence $\underset{\sim}{L}$ is a definite matrix throughout the state space.

Footnotes

1. For example, it is sufficient that the first partial derivatives of $\underset{\sim}{f}$ exist and are continuous.
2. See [1], pp. 111-112.
3. If the equilibrium position is at some other position, this may be achieved by a displacement of the coordinates.
4. Except the case that $\underset{\sim}{0}$ is quasi-asymptotically stable but not stable. In this case the system first runs away from $\underset{\sim}{0}$ and then qpproaches it asymptotically.
5. See [1], p. 66.
6. This ensures that the origin is an equilibrium position when there is no external excitation.
7. See [9], p. 176.
8. See [10], p. 126.
9. See [8], p. 303 or [9], p. 183.
10. See Eq. (10) in the Appendix. This matrix arises in the solution for $\underset{\sim}{v}_{\delta}$ and $\underset{\sim}{i}_{\gamma}$ after setting the resistor voltages and currents equal to zero in (10).
11. For example, the equation is easily solved on an analog computer.
12. See [8], p. 305.
13. See [9], p. 184.
14. See [9], p. 185.
15. See [10], p. 142.

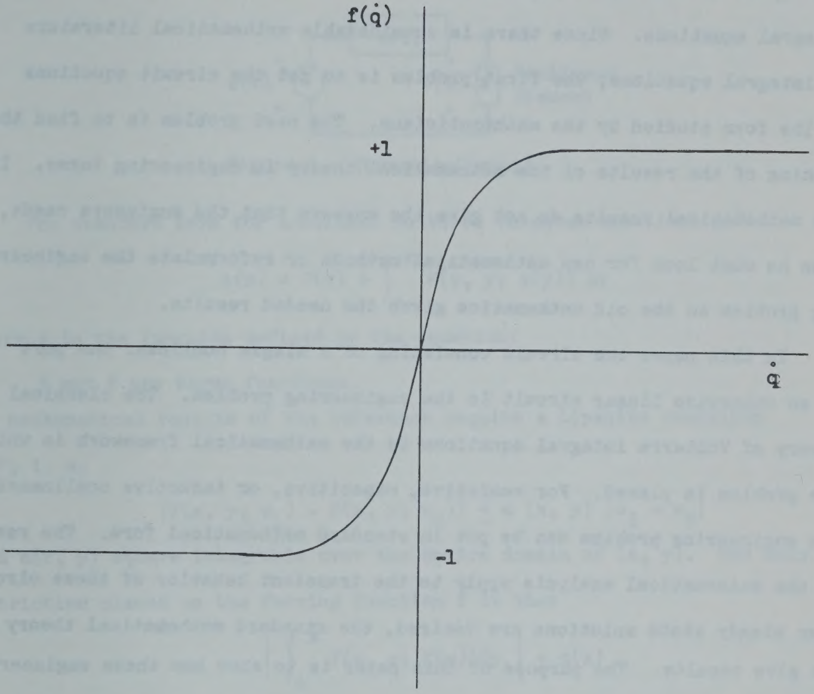


Fig. 1 Voltage-current relation for the resistive element used in the Example.

LIMITATIONS ON THE CLASSICAL VOLTERRA EQUATION THEORY IN NONLINEAR CIRCUIT PROBLEMS*

B. J. LEON
Purdue University
Lafayette, Indiana

In high order circuits containing both linear and nonlinear elements one way of formulating the physical problem mathematically is in terms of integral equations. Since there is considerable mathematical literature on integral equations, the first problem is to get the circuit equations in the form studied by the mathematicians. The next problem is to find the meaning of the results of the mathematical theory in engineering terms. If the mathematical results do not give the answers that the engineers needs, then he must look for new mathematical methods or reformulate the engineering problem so the old mathematics gives the needed results.

In this paper the circuit consisting of a single nonlinear one port in an otherwise linear circuit is the engineering problem. The classical theory of Volterra integral equations is the mathematical framework in which the problem is placed. For resistive, capacitive, or inductive nonlinearities the engineering problem can be put in standard mathematical form. The results of the mathematical analysis apply to the transient behavior of these circuits. When steady state solutions are desired, the standard mathematical theory does not give results. The purpose of this paper is to show how these engineering problems can be put into Volterra form and to point out why the classical Volterra equation theory cannot be used "as is" to get certain important engineering results.

To begin we consider a nonlinear resistance in a linear, time-invariant network with excitation. When the linear network and excitation are reduced by Thevenin's theorem the result is the one loop circuit of Fig. 1. For this circuit Kirchhoff's law gives

$$e(t) = \int_a^t z(t - \tau)\, i(\tau)\, d\tau + v(t) \tag{1}$$

---

* This research was supported by the National Science Foundation under Grant GP-581.

where e(t) is the Thevenin voltage including the effect of initial energy storage in the linear circuit at t = a, and z(t) is the driving point impulse response function of the linear circuit from the terminals of the nonlinear element.

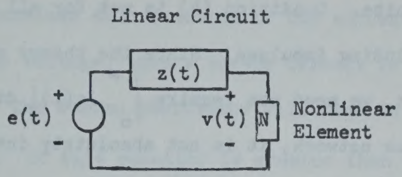


Figure 1. Thevenin Circuit

The standard form for nonlinear Volterra integral equations is[1]

$$\phi(x) = f(x) + \int_0^x F[x, y; \phi(y)]\, dy \qquad (2)$$

where ϕ is the function defined by the equation;

f and F are known functions.

The mathematical results of the reference require a Lipshitz condition on F; i. e.

$$|F(x, y; w_1) - F(x, y; w_2)| \leq a\,(x, y)\,|w_1 - w_2| \qquad (3)$$

with a(x, y) square integrable over the entire domain of (x, y). The only restriction placed on the forcing function f is that

$$\left| \int_0^x F(x, y; f(y))\, dy \right| \leq n(x) \qquad (4)$$

with n(x) square integrable.

When the above restrictions are met, eq. (2) has a unique continuous solution for any finite x.

The nonlinear resistance circuit equation (1) can readily be put in Volterra form (2), if the current through the nonlinear resistive element is a function of the voltage, and if the impulse response function z(t) has no impulses. If we use i to represent the function that gives the current in the nonlinear resistive element when the voltage across the element is known, eq. (1) can be written in Volterra form.

---

[1] F. G. Tricomi, Integral Equations, sec. 1.13, Interscience Publishers, New York, 1957

$$v(t) = e(t) - \int_a^t z(t - \tau)\, \underline{i}(v(\tau))\, d\tau \tag{5}$$

If the response function z(t) is a "good" function containing no impulses, doublets, etc., then condition (3) is met as long as the nonlinear conductance, $G(v) = \frac{d\underline{i}}{dv}(v)$, is finite. Condition (4) is met for all normal engineering excitations e(t) including impulses. Since the theory of eq. (4) leads to a solution for finite t, we need not require $\int_o^\infty |z(t)|\, dt < \infty$. If z(t) is the response of a lossless network, it is not absolutely integrable on [o, ∞], but it is square integrable on any interval [o, t].

When z(t) contains a doublet at the origin, its Laplace transform has a pole at infinity. In this case the inverse operator y(t) that would be used in a Norton equivalent circuit would be a "good" function. Now so long as the voltage can be expressed as a function of the current, the circuit dual to Fig. 1 leads to a Volterra equation like eq. (5) but with dual quantities.

When z(t) contains an impulse at the origin, the Laplace transform, Z(s), is regular at infinity. The amplitude of the impulse is Z(∞). In this case

$$\int_a^t z(t - \tau)\, i(\tau)\, d\tau = \int_a^t z_1(t - \tau)\, i(\tau)\, d\tau + Z(\infty)\, i(t) \tag{6}$$

$$\text{where } z_1(t) = L^{-1}\,[Z(s) - Z(\infty)].$$

The new operator $z_1(t)$ may not have a PR transform, but it is still square integrable for any finite t. The term Z(∞) i(t) can now be combined with the voltage across the nonlinear resistance. These two voltages can be added only if the voltage across the nonlinear resistance is uniquely determined as a function of the current.

In this last case the voltage for the nonlinear element is $\underline{v}(i)$. The total voltage of the new nonlinear element is $\underline{v}_1(i) = \underline{v}(i) + Z(\infty)i$. For the original nonlinear element there may be a region of negative resistance and then the current would not be given as a function of the voltage. However, after the positive resistance Z(∞) is added, the total voltage function must have an inverse. In other words Z(∞) must be at least as large as the maximum negative resistance. If we let $\underline{i}_1$ be the function inverse to $\underline{v}_1$, this case is characterized by the Volterra equation

$$v_1(t) = e(t) - \int_a^t z_1(t - \tau)\ \underline{i}_1(v_1(\tau))\ d\tau. \qquad (7)$$

After eq. (7) is solved for v_1 the current in the circuit is given by $i(t) = \underline{i}_1(v_1(t))$

When the impedance Z(s) is regular at infinity and the voltage across the nonlinearity is not determined as a function of the current but the current is a function of the voltage, then a Norton circuit is more appropriate than the Thevenin circuit. The Norton admittance will also be regular at infinity with $Y(\infty) = \frac{1}{Z(\infty)}$. If this quantity is greater than the maximum negative conductance of the nonlinear element, then we can get a Volterra equation like eq. (7) but with dual quantities.

In summary we have the following situations for nonlinear resistances. First if the resistance is positive, the equation can always be put in Volterra form. If there is a negative resistance region and the element is voltage controlled, we can put the equation in Volterra form if the nonlinear resistance is shunted by a capacitance or if Y(∞) is greater than the maximum negative conductance. If the nonlinear resistance is current controlled through the negative resistance region, then we must have a series inductance or have Z(∞) greater than the maximum negative resistance. Fortunately the usual circuit model for the tunnel diode, voltage control with a shunt capacitance is included in the allowable cases.

A nonlinear capacitance differs from a nonlinear resistance in that a charge - voltage relationship characterizes the capacitance as the current - voltage relationship characterizes the resistance. Figure 1 still applies as does eq. (1). This equation can be converted to a charge - voltage equation by integrating by parts. If z(t) contains no impulses or doublets, the result is

$$e(t) = z(o)\ q(t) - z(t - a)\ q(a) + \int_a^t z'(t - \tau)\ q(\tau)\ d\tau + v(t) \qquad (8)$$

If the voltage v(t) can be found as a function of the charge q(t), call the functional relation $\underline{v}$, then, eq. (8) can be regrouped as

$$\underline{v}(q(t)) + z(o)\, q(t) + \int_a^t z'\,(t - \tau)\, q(\tau)\, d\tau = e(t) + z(t - a)\, q(a) \qquad (9)$$

If we now let

$$\underline{v}_1(q) = \underline{v}(q) + z(o)$$

we get a Volterra equation if $\underline{v}_1$ has an inverse, $\underline{q}_1$. The equation is

$$v_1(t) = e(t) + z(t - a)\, q(a) - \int_a^t z'\,(t - \tau)\, \underline{q}_1(v_1(\tau))\, d\tau \qquad (10)$$

The derivation of eq. (10) carries through provided z(t) is a "good" function and $\underline{v}_1(q)$ has an inverse. If the original capacitance is positive these conditions can always be met. First, a small constant capacitance can always be separated from the nonlinear capacitance and associated with the linear network. This assures the z(t) is satisfactory. The remaining nonlinear capacitance will still be positive. For $\underline{v}_1$ to have an inverse requires $\underline{v}_1'(q) > 0$. Since the nonlinear capacitance represented by $\underline{v}$ is positive, $\underline{v}_1'$ is even more positive.

The cases of negative capacitance can be investigated in a manner similar to the negative resistance cases. However, at present there is little practical interest in negative capacitance. When the nonlinear element is an inductance the situation is dual to the capacitance case. When either the inductance or capacitance or resistance exhibits hysteresis, the equation cannot be put in Volterra form.

With the various circuit problems in standard form, the next question is: "What can the mathematical theory say about the circuit behavior?" The theory says that for any of the allowable circuits the response at any time t is a unique continuous function of the circuit parameters, the state of the circuit at some initial time a in the past, and the excitation from time a until t. The time a is arbitrary; any time at which the complete state of the linear circuit is known will do. The theory also gives a method of computing the response at time t, but there are better computational techniques than that in Tricomi.

In most communications systems applications of nonlinear elements the response of interest is essentially a steady state response. The prime examples are mixers, modulators, frequency multipliers, parametric amplifiers, and oscillators. The theory of Volterra equations develops the solution in a series form that goes essentially like $\sum_{n=0}^{\infty} K \frac{(t-a)^n}{n!}$. Thus one cannot take the limit as t approaches infinity and look at the steady state solution. Any estimate of the properties of the solution that is made will require more and more terms of the series as t increases.

One method [2] has been presented that gets around the difficulties for large t. The approach of that paper was to find the conditions wherein the series solution could be estimated independent of t. These conditions are much more restrictive than (3) and (4) in some respects, but they do allow impulses in the response function z(t). In ref. 2 a uniform estimate of the dynamic range over which the nonlinear element operates was obtained. Also a comparison between the steady state solution and the total solution at time t was obtained. This latter estimate gave a prediction of the rate of decay of transients. Neither of these engineering results could have been obtained by the classical Volterra theory.

The difficulties with the results of ref. (2) are two fold. First the conditions that replace (3) and (4) exclude cases of engineering interest if the excitation is not very small. Second the estimates obtained are quite conservative. The paper did make a good first step away from the restrictions of the Volterra theory, but there are many more steps to be taken.

---

[2] B. J. Leon and D. R. Anderson, "Stability of Pumped Nonlinear Reactance Circuits," IEEE Trans. Ct. - 10, December, 1963.

FORCED OSCILLATIONS IN NONLINEAR SYSTEMS WITH HYSTERISIS

RUI P. DE FIGUEIREDO
Purdue University
Lafayette, Indiana

1. Preliminary Remarks

The question of the existence of subharmonic oscillations in forced nonlinear systems has received considerable attention in circuit and system theory in connection with the problem of frequency conversion. The present paper deals with this question for systems possessing a general type of hysterisis property.

The phenomenon of hysterisis may perhaps be understood most readily in terms of the relationship between the current i and the flux ϕ in an iron core inductor. At any instant t_1, the value of i is not uniquely determined[1] by the value of ϕ and its derivatives at t_1 (as would be the case if i were an "ordinary" function of ϕ and its derivatives) but depends on the way ϕ varied with t before t_1. In order to take this fact into account, Block[2] expressed the relationship between two state variables like i and ϕ as an operator which uniquely maps a function space Φ, to which all possible functions $\phi(t)$ belong, into a function space I, to which all possible functions i(t) belong. In particular he considered the two-dimensional system

$$\ddot{x} + \omega_0^2 \dot{x} + \varepsilon G(x) = P(t), \quad (\dot{} = \frac{d}{dt}) \tag{1}$$

---

*This research was supported by the National Science Foundation under Grant G P - 581.

[1] In what follows it will be convenient to assume ϕ to be the independent variable although in a physical situation it is regarded as the dependent quantity.

[2] H. D. Block, "Periodic Solutions of Forced Systems having Hysterisis", IRE TRANS. ON CIRCUIT THEORY", vol. CT-7, pp. 423-431; Dec. 1960.

where $P(t)$ is, for $t \geq 0$, defined continuous and periodic with period T, ω_0 and ε are positive constants and $\omega_0 \neq \frac{2\pi n}{T}$, n = positive integer, and $G(x)$ is a hysterisis operator which, among other conditions satisfies the property[3] that if $x(t)$ is periodic with period T for $t \geq 0$ then $G(\dot{x}(t))$ is periodic with the same period for $t \geq T$. Under these conditions, he showed that there are some positive constants ε_0 and P_0 such that if $\varepsilon \leq \varepsilon_0$ and $|P(t)| \leq P_0$ $(t \geq 0)$, (1) has a unique solution which is periodic in t of period T for $t \geq T$. This solution was obtained by a process of successive iterations. However, Block's method does not apply to periodic solutions of subharmonic frequency (i.e. periodic solutions having a non-zero sinusoidal component of period nT, n = integer > 1).

Here an n-dimensional system (equation (3)) possessing the hysterisis property exhibited by (1) will be considered and it will be shown how a procedure developed by Cesari, Hale, Gambill and others[4] for periodic solutions of ordinary differential equations may be formulated for such a differential functional system. By this formulation it is possible to predict and calculate periodic solutions of subharmonic frequency as shown in the examples given in sections 3 and 4.

2. Mathematical Formulation

We shall now indicate how the theory in Chapter VI of Hale's book[4] may be adapted to the present situation. Henceforth Hale's notation will be adhered to wherever possible.

Let S_1 be the space of all n-vector functions defined continuous on $t \geq 0$, periodic of period T for $t \geq T$ and with the norm[5] $\nu(f(t)) = \sup\limits_{T \leq t \leq 2T}$

---

[3] It has been pointed out by Block[2] that a large class of physical systems possessing hysterisis embody this property. For the example of section 4 of the present paper this property is illustrated in Fig. 4b.

[4] J. K. Hale, "Oscillations in Nonlinear Systems", McGraw-Hill, New York, 1963.

[5] The symbol $\|\ \|$ in $\|f(t)\|$ stands for the vector norm of $f(t)$. Also, we emphasize that throughout this paper the norm ν is defined only for $t \geq T$.

$\| \mathfrak{f}(t) \|$, and let S_2 consist of those functions in S_1 which are periodic over $t \geq 0$ and have a continuous derivative.

For every $\mathfrak{f} \in S_1$ define the operator P_0 by

$$P_0 f = \frac{1}{T} \int_{t_0}^{t_0+T} f(t)\, dt, \quad t_0 \geq T \tag{2}$$

We shall be dealing with systems which may be reduced to the form

$$\dot{z} = \varepsilon \mathcal{F} \left\{ Z(t, z, \varepsilon) \right\} \tag{3}$$

where ε is a parameter ; z and Z are n-vectors ; $Z(t, z, \varepsilon)$ is periodic in t with period T and has continuous partial derivatives with respect to t, z and ε, for $t \geq 0$, $\| z \| \leq R$, $R > 0$, and $|\varepsilon| \leq \varepsilon_0$, $\varepsilon_0 > 0$; and $\mathcal{F}$ is an operator that maps every continuous function $Z(t, z(t), \varepsilon) = Z^*(t, \varepsilon)$ defined above into a continuous function $F(t, \varepsilon)$ on $t \geq 0$ and satisfies the following three conditions:

(a) If $Z^*(t, \varepsilon) \in S_2$ then $\mathcal{F}(Z^*(t, \varepsilon)) \in S_1$

(b) Let D consist of the functions $\mathfrak{f} \in S_1$ with $\nu(f) \leq R$,

Then there is a constant k such that for every $u(t)$ and $v(t) \in D$

$$\begin{aligned} \nu(\mathcal{F}(Z(t, u(t+T), \varepsilon)) - \mathcal{F}(Z(t, v(t+T), \varepsilon))) &\leq \\ k\, \nu(u(t+T) - v(t+T)) &= \\ k\, \nu(u(t) - v(t)) & \end{aligned} \tag{4}$$

(c) There is a $M > 0$ such that for every $z \in D$ (where D is as in (b))

$$\nu(\mathcal{F}(Z(t, z, \varepsilon))) \leq M \tag{5}$$

For a given constant n-vector a such that $\| a \| \leq b < d < R$, where b and d are constants, define the set S_0 made of the functions $\mathfrak{f} \in S_1$ for which $P_0 \mathfrak{f} = a$ and $\nu(\mathfrak{f}) \leq d$.

If $z(t) \in S_0$ introduce the operator H in the following way:

$$H F(t, \varepsilon) = a + \varepsilon \int_{\xi}^{t} (I - P_0) F(\tau, \varepsilon)\, d\tau, \quad t \geq T \tag{6a}$$

$$F(t, \varepsilon) = \mathcal{F}(Z(t, z(t+T), \varepsilon)) \tag{6b}$$

In (6a), I is the identity operator and the constant $\xi \geq T$ is chosen so that the integral is the unique periodic primitive of zero mean value for $t \geq T$.

Now from (6 a), clearly

$$P_0 \; H F(t,\varepsilon) = a \tag{7}$$

and due to the way in wich ξ was selected, it follows from (6 a), (6 b) and (5) that

$$\begin{aligned}
\| H \mathcal{F}(Z(t, z(t+T, \varepsilon))\| &\leq \|a\| + |\varepsilon| \, \Big\| \int_{\xi}^{t} (I - P_0) F(\tau, \varepsilon) d\tau \Big\| \\
&\leq \|a\| + |\varepsilon| \int_{\xi}^{\xi + T} \nu \left[(I - P_0) F(\tau, \varepsilon) \right] d\tau \\
&\leqq \|a\| + 2 |\varepsilon| \; TM \\
&\leq b + 2 \; |\varepsilon| \; TM
\end{aligned} \tag{8}$$

Let

$$\varepsilon_1 = \min \left\{ \frac{d-b}{4TM} , \varepsilon_0 \right\} \tag{9}$$

Then for $|\varepsilon| \leq \varepsilon_1$, we have from (8)

$$\| H \; \mathcal{F}(Z(t, z(t+T), \varepsilon) \| \leq b + \frac{d-b}{2} < d \tag{10}$$

and hence by (7) and (10) $H \mathcal{F}(Z(t, z(t+T), \varepsilon) \in S_0$.

Moreover for every $u(t)$ and $v(t) \in S_0$, in view of (4) ,

$$\begin{aligned}
\| H \mathcal{F}(Z(t, u(t+T), \varepsilon) - H \mathcal{F}(Z(t, v(t+T), \varepsilon)) \| &\leq \\
&\leq 2 \; |\varepsilon| \int_{\xi}^{\xi + T} \| F_{(u)}(\tau, \varepsilon) - F_{(v)}(\tau, \varepsilon) \| \, d\tau \\
&\leq 2 \; |\varepsilon| \, T \; \nu \left[\mathcal{F}(Z(t, u(t+T), \varepsilon) - \mathcal{F}(Z(t, v(t+T), \varepsilon)) \right] \\
&\leq 2 \; |\varepsilon| \, T \, k \; \nu \left[u(t+T) - v(t+T) \right] \\
&= 2 \; |\varepsilon| \, T \, k \; \nu \left[u(t) - v(t) \right]
\end{aligned} \tag{11}$$

If

$$\varepsilon_2 = \min \left\{ \varepsilon_1 , \frac{1}{4Tk} \right\} \tag{12}$$

by (11), $H \mathcal{F}(Z(t, z(t+T), \varepsilon)) \subset S_0$ is a contraction for $|\varepsilon| \leq \varepsilon_2$ and therefore has a unique fixed point in S_0 .

It is clear from (6 a) and (6 b) that this fixed point is , for $t \geq T$, a solution of the system.

$$\begin{aligned}
\dot{z} &= \varepsilon (I - P_0) \; \mathcal{F}(Z(t, z(t+T), \varepsilon)) \\
&= \varepsilon (I - P_0) \; \mathcal{F}(Z(t, z(t), \varepsilon))
\end{aligned} \tag{13}$$

since $z(t+T) = z(t)$ for $t \geq T$.

If in addition we require that

$$P_0 \mathcal{F}(Z(t, z(t), \varepsilon)) = 0 \tag{14}$$

$z(t)$ is, for $t \geq T$, a periodic solution of (3) of period T. The n algebraic equations (14) constitute the so called "determining equations" of the system.

It follows[6] that this solution z(t) of (3), for $t \geq T$, is obtained as the uniform limit of the sequence $\left\{ z^{(k)}(t) \right\}$ where

$$\begin{aligned} z^{(0)}(t) &= a \\ z^{(k+1)}(t) &= a + \varepsilon \int_{\xi}^{t} (I - P_0) F^{(k+1)}(\tau, \varepsilon) d\tau, \quad \xi, t \geq T, \; k = 0, 1, 2, \ldots \\ F^{(k+1)}(t, \varepsilon) &= \mathcal{F}(Z(t, z^{(k)}(t+T), \varepsilon)), \quad k = 0, 1, 2, \ldots \end{aligned} \tag{15}$$

where at each $(k+1)^{th}$ iterative step we require that

$$P_0 \mathcal{F}(Z(t, z^{(k+1)}(t+T), \varepsilon)) = 0 \tag{16}$$

These are the n "approximate (truncated) determining equations" associated with the $(k+1)^{th}$ iteration. Although any desired degree of accuracy may be secured by means of a sufficient number of iterative steps, we shall not go beyond the first iteration in the examples that follow.

In treating the case of subharmonic oscillations, a suitable transformation is introduced which converts the original system to the form (3). By means of this transformation, z is equated to the amplitude of the subharmonic component and the period T to T' = the period of this component. This is illustrated in the two examples discussed below.

3. Third Order Subharmonic Solution in a Piecewise Linear System Without Hysterisis

As a prelude to the example in the following section we shall consider here the existence, and obtain an estimate of the amplitude, of the third order subharmonic in a piecewise linear system without hysterisis excited by a sinusoidal forcing function of t. Although this case (without hysterisis) could be treated as an ordinary differential equation, we shall use the formalism just described for the purpose of illustration. Accordingly, we shall assume that the system is defined only for $t \geq 0$ and periodicity will be investigated over $t \geq T'$.

---

[6] This is clear from (11) and (12).

Our system is a lossless L - C circuit excited by a voltage source $E_m \sin \omega t$, which yields the equation

$$N \dot{\phi} + \frac{1}{C} \int i\,dt = E_m \sin \omega t \tag{17}$$

where i and $N\phi$ are the current and flux linkage in the inductor.

We assume that, for $t \geq T' = \frac{2\pi}{\omega'} = \frac{6\pi}{\omega}$, the i vs ϕ relationship is given by

$$\frac{1}{NC}\, i(\phi) = G(\phi) = \omega^2 \phi + \varepsilon\, G_1(\phi) \tag{18}$$

$$G_1(\phi) = \begin{cases} 0, & |\phi| \leq c \\ m(\phi - c), & \phi \geq c \\ m(\phi + c), & \phi \leq -c \end{cases} \tag{19}$$

where ω^2, ε, m and c are positive constants. $G(\phi)$ is shown in Fig. 1.

(17) and (18) lead to

$$\dot{\phi} = y, \quad \dot{y} = -\omega_0^2 \phi - \varepsilon\, G_1(\phi) + A \cos \omega t \tag{20}$$

where

$$A = \frac{E_m \omega}{N} \tag{21}$$

By means of the following transformation given by Hale[7] for Duffing's equation

$$\phi = z_1 \sin \tfrac{1}{3} \omega t + z_2 \cos \tfrac{1}{3} \omega t + B \cos \omega t \tag{22a}$$

$$y = \tfrac{1}{3} \omega \left(z_1 \cos \tfrac{1}{3} \omega t - z_2 \sin \tfrac{1}{3} \omega t \right) - B\omega \sin \omega t \tag{22b}$$

where

$$B = \frac{A}{\omega_0^2 - \omega^2} \tag{23}$$

the system (20) becomes

---

[7] See reference given in footnote 4 , p. 51.

$$\dot{z}_1 = -\frac{3\varepsilon}{\omega}\Big[\beta\,(z_1 \sin\tfrac{1}{3}\,\omega t + z_2 \cos\tfrac{1}{3}\,\omega t) + G_1(z_1 \sin\tfrac{1}{3}\,\omega t + z_2 \cos\tfrac{1}{3}\,\omega t + B\cos\omega t)\Big]\cos\tfrac{1}{3}\,\omega t \qquad (24a)$$

$$\dot{z}_2 = \frac{3\varepsilon}{\omega}\Big[\beta\,(z_1 \sin\tfrac{1}{3}\,\omega t + z_2 \cos\tfrac{1}{3}\,\omega t) + G_1(z_1 \sin\tfrac{1}{3}\,\omega t + z_2 \cos\tfrac{1}{3}\,\omega t + B\cos\omega t)\Big]\sin\tfrac{1}{3}\,\omega t \qquad (24b)$$

where

$$\beta = \frac{1}{\varepsilon}\,(\omega_0^2 - \frac{1}{9}\,\omega^2) \qquad (25)$$

Letting

$$Z(t,z) = (Z_1(t,z_1,z_2),\ Z_2(t,z_1,z_2)) \qquad (26)$$

where

$$Z_1(t,z_1,z_2) = z_1 \sin\tfrac{1}{3}\,\omega t + z_2 \cos\tfrac{1}{3}\,\omega t \qquad (27)$$

$$Z_2(t,z_1,z_2) = Z_1\ (t,z_1,z_2) + B\cos\omega t \qquad (28)$$

and substituting (27) and (28) into (24 a) and (24 b) we obtain

$$\dot{z}_1 = -\frac{3\varepsilon}{\omega}\Big[\beta Z_1(t,z_1,z_2) + G_1\,(Z_2(t,z_1,z_2))\Big]\cos\tfrac{1}{3}\,\omega t \qquad (29a)$$

$$\dot{z}_2 = \frac{3\varepsilon}{\omega}\Big[\beta Z_1(t,z_1,z_2) + G_1\,(Z_2(t,z_1,z_2))\Big]\sin\tfrac{1}{3}\,\omega t \qquad (29b)$$

This set of equations , when written as a single vector equation, is identified with (3). Clearly $\mathcal{T}$ and z in the present case meet the conditions stated in the preceeding section, and so , for sufficiently small $|\varepsilon|$, we may use the procedure given in that section.

At each iteration there are two approximate determining equations[8] to be solved corresponding to (29 a) and (29 b) respectively.

Let $z^{(0)} = (b_1, b_2)$, where b_1 and b_2 are arbitrary constants.

Then the first set of approximate determining equations is :

$$\int_{T'}^{2T'}\Big[\beta Z_1(t,b_1,b_2) + G_1\,(Z_2(t,b_1,b_2))\Big]\cos\tfrac{1}{3}\,\omega t\,dt = 0 \qquad (30a)$$

$$\int_{T'}^{2T'}\Big[\beta Z_1(t,b_1,b_2) + G_1\,(Z_2(t,b_1,b_2))\Big]\sin\tfrac{1}{3}\,\omega t\,dt = 0 \qquad (30b)$$

Let $b = (b_1^2 + b_2^2)^{1/2}$ and $\alpha = 3\sin^{-1}\frac{b_1}{b}$.

For ease in integration of (30 a) and (30 b) we shall only consider the case in which

$$b^2 > 9\,|B|^2 + (\,|B| + c)^2 \qquad (31)$$

---

[8] Since (18) is defined only for $t \geq T'$ we do not have available the symmetry property of the system with respect to $t = 0$ axis which Hale[4] used to eliminate one of the determining equations for Duffing's equation.

Then, with the substitution $\psi = \frac{1}{3}(\omega t+\alpha)$, the argument of G_1 in (24) becomes

$$f(\psi) = b\cos\psi + B\cos(3\psi+\alpha) \tag{32}$$

so that by (31), letting $\psi_1 = \cos^{-1}\frac{|B|+c}{|b|}$ and assuming[9] ,$b > 0$, we have $f(\psi) > c$ for $0 \le \psi \le \psi_1$ and $2\pi - \psi_1 \le \psi \le 2\pi$; $f(\psi) < -c$ for $\pi - \psi_1 \le \psi \le \pi + \psi_1$; $f'(\psi) < 0$ for $\psi_1 \le \psi \le \pi - \psi_1$; and $f'(\psi) > 0$ for $\pi + \psi_1 \le \psi \le 2\pi - \psi_1$. Hence in a single (subharmonic) period, the graph of $y = f(\psi)$ intersects the graph of $y = c$ at two and only two points, say a $t = t_1$ and t_4 (see Fig. 2). Similarly in that period it intersects the graph of $y = -c$ at two and only two points, say at t_2 and t_3 . Note also that $t_3 = t_1 + \frac{1}{2}T'$ and $t_4 = t_2 + \frac{1}{2}T'$ since $f(\psi)$ contains only odd harmonics.

It is thus possible to split up the interval $(T', 2T')$ into only five subintervals separated by points t_i , $i = 1, 2, 3; 4$, and calculate the two integrals (30 a) and (30 b) using the relationships (19), (27), (28) and (29). The final result is

$$\left[\frac{4\pi}{\varepsilon m}\left(\omega_0^2 - \frac{\omega^2}{9}\right) + 4(\pi + \theta_1 - \theta_2)\right] b_2 + 2(\sin 2\theta_1 - \sin 2\theta_2)(b_2 + B)$$
$$+ (\sin 4\theta_1 - \sin 4\theta_2)B - 2(\cos 2\theta_1 - \cos 2\theta_2)b_1 - 8c(\sin\theta_1 + \sin\theta_2) = 0 \tag{33a}$$

$$\left[\frac{4\pi}{\varepsilon m}\left(\omega_0^2 - \frac{\omega^2}{9}\right) + 4(\pi + \theta_1 - \theta_2) - 2(\sin 2\theta_1 - \sin 2\theta_2)\right] b_1 +$$
$$(-\cos 4\theta_1 + \cos 4\theta_2)B + 2(-\cos 2\theta_1 + \cos 2\theta_2)(b_2 - B) + 8c(\cos\theta_1 + \cos\theta_2) = 0 \tag{33b}$$

where

$$\theta_1 = \frac{1}{3}\omega t_1 \tag{34a}$$

$$\theta_2 = \frac{1}{3}\omega t_2 \tag{34b}$$

---

[9] Similar considerations apply if $b < 0$.

In the special case[10] in which $b_1 = 0$, the angle $\theta_1 = \pi - \theta_2 = \theta$ satisfies the second determining equation (33 b) identically for arbitrary b_2, B and c. Then this equation may be ignored and (33 a) simplifies to :

$$\omega_0^2 b' = \omega_0^2 b_2' = \frac{4c'\sin\theta - (\sin 2\theta + \frac{1}{2}\sin 4\theta)\dfrac{1}{[1-(\frac{\omega}{\omega_0})^2]}}{2\theta + \sin 2\theta + \dfrac{\pi\omega_0^2}{9\varepsilon m}\left[9 - (\frac{\omega}{\omega_0})^2\right]} \tag{35}$$

where the primed quantities are the imprimed quantities divided by A.

For $c' = .05$ and $\frac{\varepsilon m}{\omega_0^2} = .1$, and for values of $\theta, 0 \leq \theta \leq \frac{\pi}{2}$, satisfying

$$b_2' \cos\theta + B' \cos 3\theta = c' \tag{36}$$

where $B' = \frac{1}{A} B$, the curve labeled Λ in Fig. 7 shows the plot of $\omega_0^2 b'$ versus $(\frac{\omega}{\omega_0})^2$. The approximation (31) is valid on the portion of this curve above the point M.

4. Third Order Subharmonic Oscillation in a Piecewise Linear System With Hysterisis

Let us now consider the case in which the i versus ϕ relationship exhibits a hysterisis loop and the periodic forcing term is of large enough amplitude to make the response span beyond the loop. This is only a simple example and other more complicated cases satisfying the hysterisis property mentioned earlier can be treated in much the same way.

The circuit is the same as before. G (ϕ) is given by (18), but instead of (19) we now have[11]

---

[10] This of course corresponds to the case in which the initial value (at $t = T'$) of the third subharmonic has maximum magnitude (see Fig. 2).

$$G_1(\phi) = \begin{cases} \text{for decreasing } \phi \begin{cases} -k_1 + m\phi\,, & (\phi \geq c_1) \\ -k_2\,, & (-c_2 \leq \phi \leq c_1) \\ k_1 + m\phi\,, & (\phi \leq -c_2) \end{cases} \\ \\ \text{for increasing } \phi \begin{cases} k_1 + m\phi\,, & (\phi \leq -c_1) \\ k_2\,, & (-c_1 \leq \phi \leq c_2) \\ -k_1 + m\phi\,, & (\phi \geq c_2) \end{cases} \end{cases} \qquad (37)$$

where

$$k_1 = \frac{1}{2} m\,(c_1 + c_2) \qquad (38a)$$

$$k_2 = \frac{1}{2} m\,(c_2 - c_1) \qquad (38b)$$

and $c_2 > c_1$ and m are positive constants.

$G(\phi)$, $\phi(t)$ and $G_1(\phi(t))$ are illustrated in Figs. 4a and b , 5 and 6.

As before, we obtain the determining equation (30 a) and (30 b).

For the case in which

$$b^2 > 9\,|B|^2 + (\,|B| + c_2)^2 \qquad (39)$$

we derive in a way similar to that used in obtaining (33 a) and (33 b):

$$\left[\frac{4\pi}{\varepsilon m}(\omega_0^2 - \frac{\omega^2}{9}) + 4(\pi + \theta_1 - \theta_2)\right] b_2 + 2(\sin 2\theta_1 - \sin 2\theta_2)(b_2 + B)$$

$$+ (\sin 4\theta_1 - \sin 4\theta_2)B - 2(\cos 2\theta_1 - \cos 2\theta_2)b_1 - 8\,(c_1 \sin\theta_1 + c_2 \sin\theta_2) = 0 \qquad (40a)$$

$$\left[\frac{4\pi}{\varepsilon m}(\omega_0^2 - \frac{\omega^2}{9}) + 4(\pi + \theta_1 - \theta_2) - 2(\sin 2\theta_1 - 2\sin 2\theta_2)\right] b_1$$

$$+ (-4\cos\theta_1 + 4\cos\theta_2)B + 2(-\cos 2\theta_1 + 2\cos 2\theta_2)(b_2 - B) + 8(c_1\cos\theta_1 + c_2\cos\theta_2) = 0$$

where θ_1 and θ_2 are as in (34).

If $b_1 = 0$ and c_1 and c_2 are small or nearly equal, the second equation nearly vanishes identically for $\theta_1 = \pi - \theta_2 = \theta$ and then (40 a) gives:

---

[11] This mapping applies only for large variations in ϕ about $\phi = 0$ (See Fig. 4a). Otherwise situations similar to those illustrated in Fig. 4b may arise.

$$\omega_0^2 b' \cong \omega_0^2\, b_2' \cong \frac{2(c_1'+c_2')\sin\theta - (\sin 2\theta + \frac{1}{2}\sin 4\theta)\,\frac{1}{\left[1-(\frac{\omega}{\omega_0})^2\right]}}{2\theta + \sin 2\theta + \frac{\pi\omega_0^2}{9\varepsilon m}\left[9 - (\frac{\omega}{\omega_0})^2\right]} \tag{41}$$

where again the primed quantities are the imprimed ones divided by A.

For $c_1' = .03$, $c_2' = .05$ and $\frac{\varepsilon m}{\omega_0^2} = .1$, and for values of θ which when substituted in the left hand side of (36) equal[12] to $\frac{1}{2}(c_1'+c_2')$ equation (41) yields the curve Γ in Fig. 7, where N is the point above which the approximation (39) holds.

---

[12] For c_1' and c_2' small θ is insensitive to their magnitude.

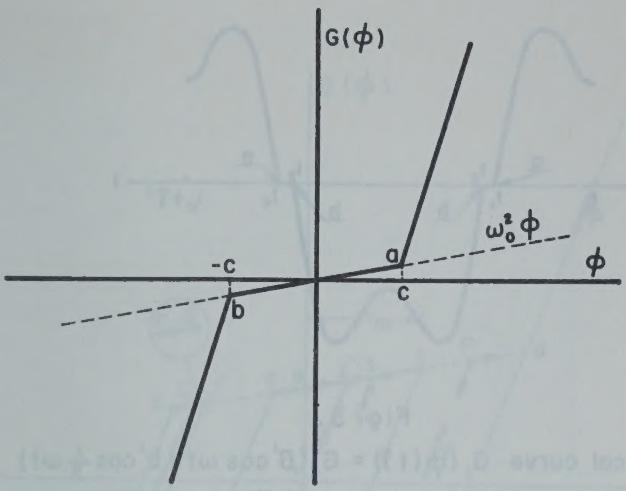


Fig. 1.

$G(\phi)$ for a system with saturation (without hysterisis)
(Note: $G_1(\phi) = \frac{1}{\epsilon}\left[G(\phi) - \omega_0^2 \phi\right]$)

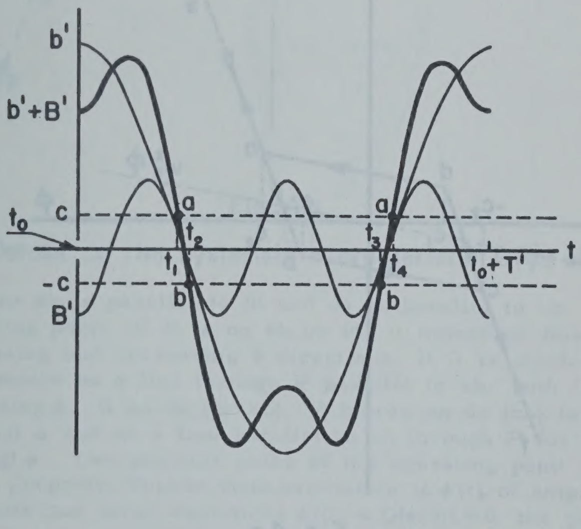


Fig. 2.

The curve $B' \cos \omega t + b' \cos \frac{1}{3} \omega t$ for the system of Fig. 1.

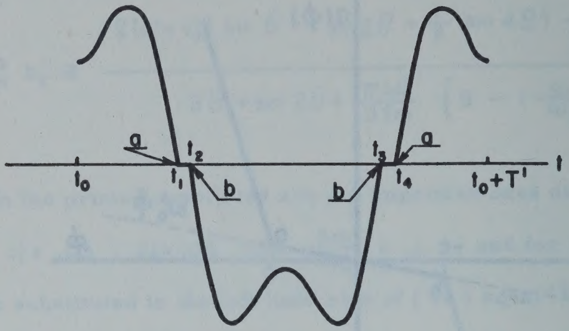


Fig. 3.

A typical curve $G_1(\phi(t)) = G_1(B'\cos\omega t + b'\cos\frac{1}{3}\omega t)$ for the system of Fig. 1.

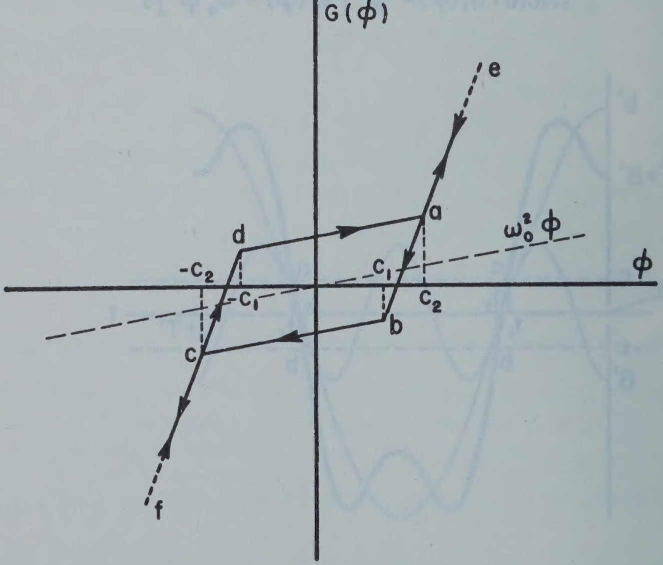

G(ϕ)
e
a
d
$\omega_0^2\phi$
$-c_2$
c_1
ϕ
$-c_1$
c_2
b
c
f

Fig. 4a.

G(ϕ) for a system with saturation and hysterisis when the amplitude of oscillation of ϕ is large. (Note: $G_1(\phi) = \frac{1}{\epsilon}\left[G(\phi) - \omega_0^2\phi\right]$)

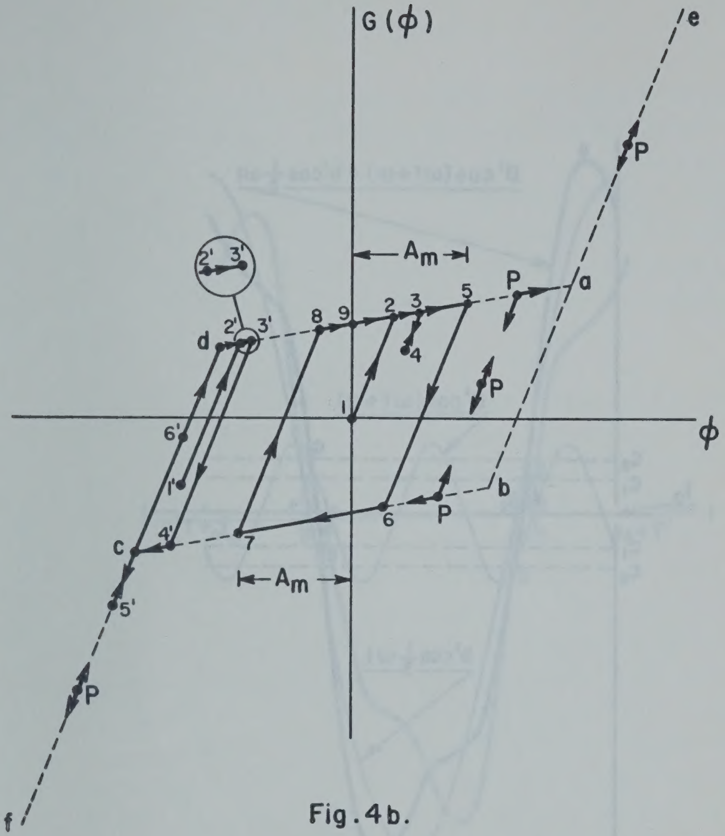


Fig. 4b.

Details on the hysterisis characteristic $G(\phi)$ of Fig. 4a

The line eb is parallel to fd and da is parallel to cb. Let P denote the operating point. If P is on eb or fd, it moves on this line both in the increasing and decreasing ϕ directions. If it is inside the trapezoid abcd, it moves on a line through P parallel to eb, both for increasing and dcreasing ϕ. If on da (or cb), P moves on da (cb) for increasing (decreasing) ϕ and on a line parallel to eb through P for decreasing (increasing) ϕ. Two possible paths of the operating point illustrate the hysterisis property. Thus for some oscillation in $\phi(t)$, of amplitude A_m, period T, with the initial conditions $\phi(0) = G(\phi(0)) = 0$, the path 1234356789 is traced in the interval $0 \leq t \leq T$. Note that beyond $t = T$, the path 9234356789 becomes periodic. Similar remarks apply to the path 1'2'3'4'c5'c6'd2'3'4'... which is periodic from 6' on. (N.B.: The point 6' lies on fd above 1').

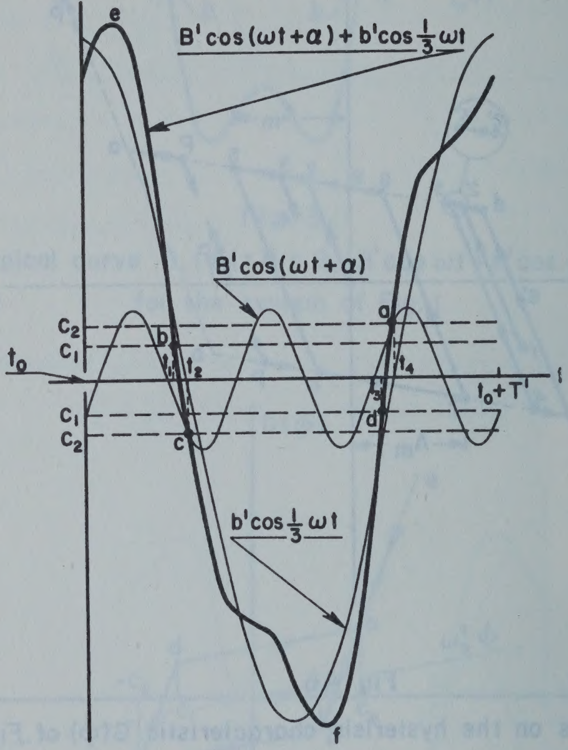


Fig. 5.

The curve $B' \cos(\omega t+\alpha)+b' \cos \frac{1}{3}\,\omega t$ for the system of Fig. 4.

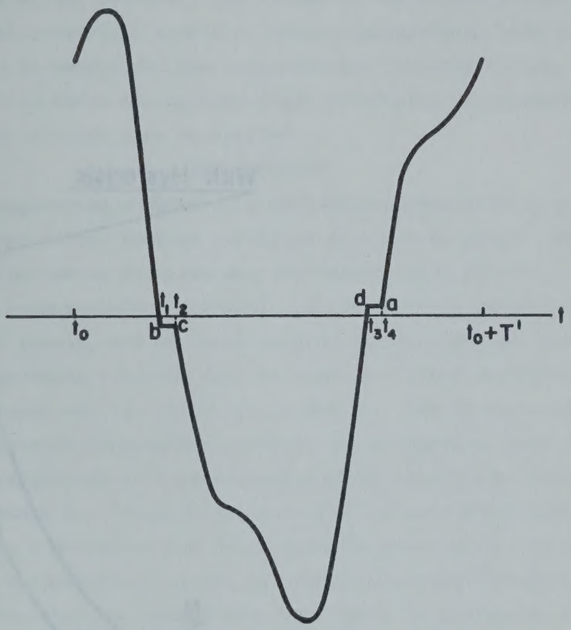

$t_1\,t_2$
d
a
t
t_0
b
c
$t_3\,t_4$
t_0+T'

Fig. 6.

A typical curve $G_1(\phi(t)) = G_1(B'\cos(\omega t+\alpha) + b'\cos\frac{1}{3}\omega t)$ for the system of Fig. 4.

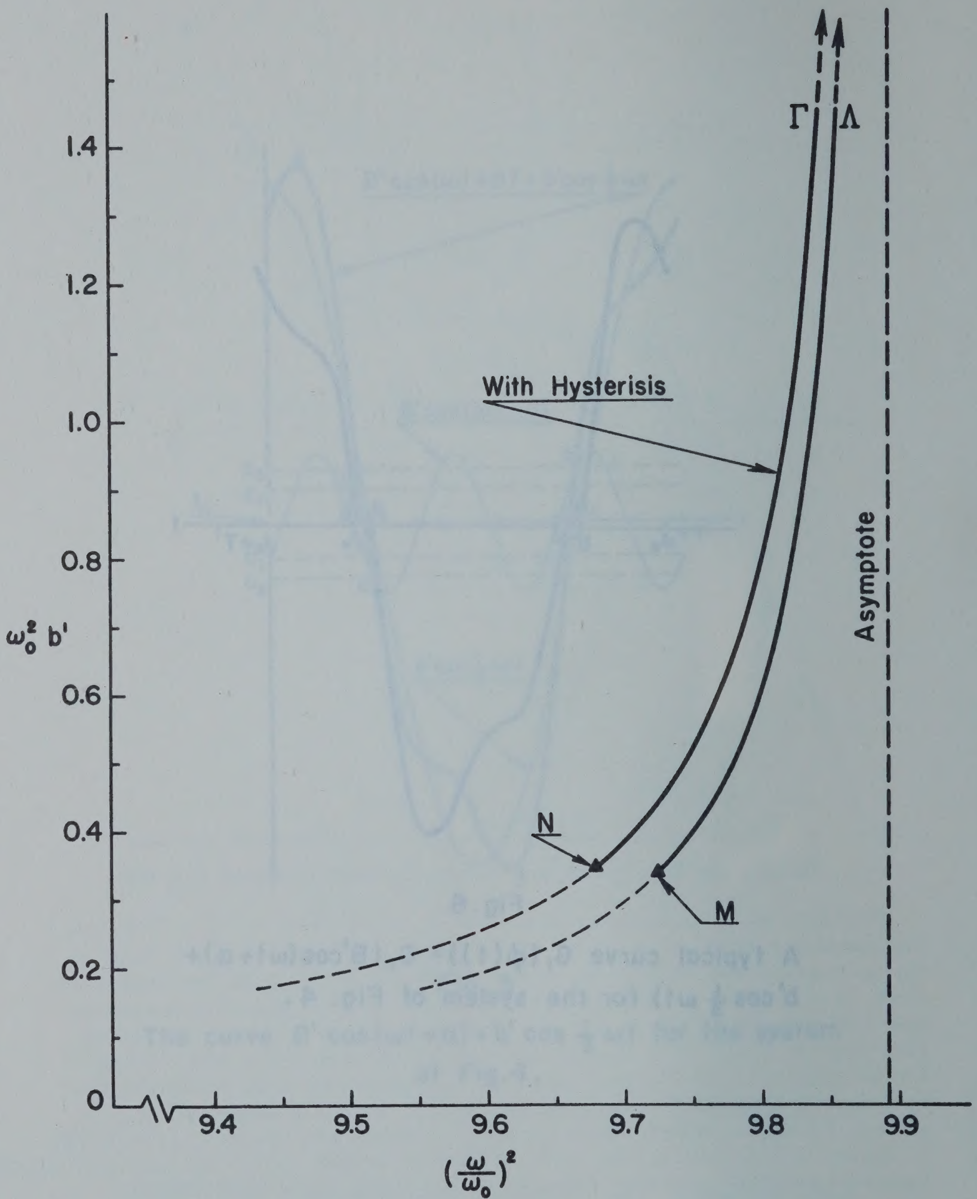


Fig. 7.

Third subharmonic frequency response curves Λ and Γ for the systems referred in Figs. 1 and 4 respectively.

DISTRIBUTED SOLUTION OF NETWORK PROGRAMMING PROBLEMS

JACK B. DENNIS
Department of Electrical Engineering and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Summary - A method is presented for the solution of network programming problems through the iterative execution of the same algorithm for each station of the network. At a given station, the algorithm requires data only from adjacent stations of the network. The theory of the method is developed through the use of electrical models of network programming problems. The techniques should be useful for the communications message-routing problem and similar cases in which the disadvantages of relying on a computation performed at a master station must be avoided.

INTRODUCTION

A network programming problem is a mathematical programming problem in which the structure of the problem statement is a linear graph. Examples of linear network programming problems are the network flow problem,[1] which incorporates the transportation problem[2] and its variants as well as a number of other familiar models, and the more general multi-commodity flow problem.[3] These linear programming examples will be used as a basis for developing the material of the paper and are explained in detail. The linear graphs associated with network programming problems often represent geographic networks of communications or transportation paths connecting commercial centers. By a method for "distributed solution" of network programming problems,[4] we mean a procedure for their solution carried out through computation performed cooperatively at the geographically distributed centers of the network. The motivation for considering this class of method is discussed following the explanation of the network flow problem in the next section.

THE NETWORK FLOW PROBLEM

Figure 1 shows a linear-directed graph that consists of a number of stations represented by dots, connected by many directed links represented by arcs with arrows. The network flow problem is a special linear programming formulation concerning flows associated with the links of a graph. Each link of the network is characterized by a cost $\underline{c}$ per unit flow and an upper bound $\underline{m}$ on the amount of flow $\underline{x}$, as indicated in Fig. 2. Only positive flows in the direction of the link are permitted. A source or sink of flow $\underline{f}$ may be associated with each station of the network. At each station flow is required to satisfy a continuity law, that is, the algebraic sum of flows into the station must be zero. A pattern of flow that meets the conditions just stated is called a feasible flow. The object of the problem is to find a feasible flow that minimizes the total cost ϕ given by the product of $\underline{c}$ and $\underline{x}$ summed over all branches.

To state the network flow problem mathematically, let the branches of the network be numbered arbitrarily, and similarly for the nodes. Let g =O(i), and h = T(i) be the numbers of the two nodes on which the branch i originates and terminates, respectively. With this notation the network flow problem is as shown below.

Minimize

$$\phi = \sum_i c_i x_i \qquad \text{(total cost)} \qquad (1)$$

with

$$0 \leqslant x_i \leqslant m_i \qquad \text{(bounds)} \qquad (2)$$

and

$$\sum_{i 3 O(i)=j} x_i - \sum_{i 3 T(i)=j} x_i = f_j \qquad \text{(continuity)} \qquad (3)$$

This is a special form of linear programming problem and a number of algorithms for its solution are known.[5-7]

The application of the network flow model to optimal distribution of goods through a transportation network is clear. The multi-commodity flow problem is a generalization of the network flow problem to apply to the distribution of diverse commodities through a common facility. The multi-commocity flow problem also describes the routing of messages in a communications network.[8]

Consider the application of the network flow model to the distribution of goods from many sources to many consumers through a complex network of routes and transshipment points covering a wide geographic region. One method of operating such a distribution system is to transmit the data of the network flow problem to a central computing facility that computes the least-cost traffic pattern and transmits appropriate portions of the solution to each station of the network. The method of distributed solution, in the sense of this paper, computes a solution through a simple computation performed iteratively at each station of the network. Each station receives information directly from only those stations that are adjacent to it in the network. As we shall see, this adjacent station information contains all data about the entire network which are necessary to perform the correct calculation. The method is an iterative procedure for computing successively more accurate solutions to the programming problem and can be used in the presence of varying data. The operational advantages of the method in lessened vulnerability to failure, reduced data transmission, and quicker response to changing conditions are evident.

THE DUAL FORMULATION

The method of distributed solution is most readily explained in connection with the electrical model of the network programming problem[9] because the analysis methods of electric circuit theory are convenient for demonstrating properties of the method and giving insight into its operation. Since the duality concept of mathematical programming is important for understanding the formulation of the electrical model, we shall review briefly the duality principle of linear programming. The primal and dual formulations of the general linear programming problem are set down in Fig. 3.‡ The theory of linear programming[10] demonstrates that: 1) An optimal solution (x,v) of the primal form, and an optimal solution (y,u) of the dual form together satisfy the complementary slackness conditions as an automatic consequence. 2) A set of values (x,v,y,u) that satisfies the constraint relations of both formulations, and satisfies the complementary slackness conditions, automatically constitutes optimal solutions for both formulations. Thus satisfaction of these sets of relations (which will be called the Lagrangian formulation of the programming problem) is equivalent to either of the extremal formulations.

The dual and Lagrangian formulations for the network flow problem may be found by relating its mathematical statement (relations (1), (2), and (3) to the general primal form, and by using Fig. 3 as a model. The results are exhibited as Fig. 4. The quantities y_g in the dual formulation may be regarded as incremental costs or prices. In the solution of the problem a difference $y_g - y_h$ is the unit cost of increasing flow within the network from station g to station h over the most favorable route.

ELECTRICAL MODEL OF THE NETWORK FLOW PROBLEM

The electrical model is constructed from the circuit elements defined mathematically in Fig. 5-ideal diodes and fixed voltage and current sources. The polarity convention used is such that the product of voltage and current is always the power delivered to the network by the circuit element.

A link of the flow network establishes a price-versus-flow relationship as shown in Fig. 6a. The unit price of flow is c_i for flows within the bounds of the link, but the cost of establishing a flow that is negative or exceeds the capacity of the link is infinite. In Fig. 6b we show an electrical network with a voltage/current relationship identical to the price/flow relation of the flow network link. Thus it appears that this electrical model of a link could form the basis of a complete electrical model of the network flow problem. This is indeed the case, and the complete model is shown in Fig. 7a. Currents in the model are analogous to flow quantities, and voltages correspond to prices. Each link of the flow network is replaced by the electrical model of a link with appropriate source values. Sources and sinks of flow are replaced by current sources of corresponding value.

The solution of an electrical network is a set of currents and voltages that satisfy

1) the Kirchhoff loop law,

2) the Kirchhoff node law,

3) the constraints imposed by circuit elements.

These principles applied to the flow network model yield the set of relations exhibited in Fig. 7b. The constraints imposed by voltage and current sources are incorporated in the Kirchhoff law equations. The relations that must be satisfied by any solution of the electrical model are seen to be precisely the statement of the Lagrangian formulation of the network flow problem. We therefore have the following principles:

1. Any solution of the electrical model provides optimal solutions of the primal and dual network flow programs.

2. Any pair of optimal solutions of the programs constitutes a solution of the electrical model.

It is interesting to note that, in terms of the electrical model, the objective function of the primal network flow program is the power absorbed from the network by the voltage sources, and the objective function of the dual is the power delivered to the network by the current sources.

DYNAMIC ELECTRICAL MODEL

The electrical model of the network flow problem is a static circuit in the sense that its present response is dependent only on present excitation, and is independent of past excitation. The circuit has no memory. Currents and voltages in the model would immediately assume values corresponding to the solution of the flow problem if the network could be assembled with ideal physical elements. If the sources in the model change value in correspondence with changing data values for the network flow problem, the currents and voltages representing flows and prices would instantly assume values corresponding to the new optimal flow pattern.

The distributed algorithm is based on simulating the behavior of the electrical model. Simulation of a static circuit by a procedure executed at each station with data received from adjacent stations is not attractive: A change in source value at one point in the static electrical model may drastically alter the optimal flow pattern at any point in the network.

The difficulty may be circumvented by altering the electrical model so that the following properties are realized:

a) With source values fixed, the solution of the altered model converges to the exact solution of the network flow problem with increasing time.

b) A change in a source value associated with a change in data has an immediate effect only on variables associated with immediately adjacent stations and links.

The second property may be realized by modifying the static model through the addition of inductors and capacitors as shown in Fig. 8. With this arrangement, a sudden jump in the current entering any node is converted into a slowly rising voltage by the capacitor there. A sudden drop in voltage across a link becomes a gradually changing current through the integrating effect of the series inductor. It is evident that a solution of the modified electrical model which is not changing with time must be a solution of the static model, since the inductor voltages and capacitor currents must all be zero. This fact by itself does not insure convergence according to the first property stated above: There are no resistive elements in the model to dissipate stored energy in the network, and transient solutions would not decay. The addition of resistances in series with the capacitors and in parallel with the inductors as shown in Fig. 9 will cause transient energy to be dissipated. This circuit will be called the dynamic electrical model of the network flow problem.

Since there is zero power dissipation in the resistances for a steady-state solution of the dynamic model, such a solution still solves the static model and therefore indicates the optimum flow pattern. Intuitively, a solution of the dynamic model should converge to a steady-state solution from arbitrary initial states of the inductors and capacitors. We shall give arguments that this is so.

If the dynamic electrical model were released from some arbitrary initial conditions, its solution would follow a course during which diodes would change state at certain instants. For any segment of time during which no diode changes state, all diodes of the network may be replaced by short or open circuits according to their state. Under these conditions the network is simply a linear RLC network with fixed sources, and possibly has an oscillatory solution of exponentially decreasing magnitude. A complete solution of the dynamic model is accordingly a "piecewise exponentially decaying" function of time. Bounds on the range of natural frequencies associated with the segments has bearing on the success of the distributed algorithm and will be duscussed below.

THE DISTRIBUTED ALGORITHM

The set of relations describing the solution of the dynamic model as a function of time τ are given in Fig. 10 in terms of the notation of Fig. 9. The integral description of the energy storage elements is used to facilitate the application of an integration rule. From these relations we obtain a set of relations for assigning values to the problem variables for time $\tau + \delta$ in terms of their values at time τ. For the present discussion the simple rectangular integration rule

$$\int_{\tau}^{\tau+\delta} \phi(\tau)\, d\tau \approx \delta\phi(\tau)$$

will be employed, although the use of the trapezoidal rule would probably be preferred in this application to reduce truncation error. The descriptions of the capacitors and inductors of the dynamic model are consequently replaced by

$$t_i(\tau+\delta) = t_i(\tau) + \frac{\delta}{\lambda} s_i(\tau), \text{ all } i$$

$$p_i(\tau+\delta) = p_i(\tau) + \frac{\delta}{K} n_i(\tau), \text{ all } i$$

$$r_g(\tau+\delta) = r_g(\tau) + \frac{\delta}{K} q_g(\tau), \text{ all } g.$$

The result of the substitution above is a system of difference equations which may be applied recursively to compute an approximate solution of the dynamic model as a function of time. The order in which the relations of Fig. 10 are used in the computation is dependent on the states of the diodes. The computation rules may be grouped by station by associating all relations for a link with its station of origin. Then the computation steps required at any particular station, say station g, are as set forth in Fig. 11. Here a_i and b_i are indicators showing the states of the diodes in link i of the model. A value of C means "closed" and 0 means "open".

CONVERGENCE OF THE SOLUTION OF THE DYNAMIC MODEL

The convergence of the solution of the dynamic model to a solution of the static model will be shown here. The arguments used are general and apply to any network constructed from fixed sources, diodes, inductors, capacitors and resistors. Consider the dynamic model in the form shown in Fig. 12a. All of the passive, linear components have been placed inside the box while the current sources, voltage sources, and diodes have been brought outside for special consideration. Let $[\overset{*}{e}(\tau),\overset{*}{i}(\tau)]$ be a set of branch voltages and currents that form a solution in time of this network. In Fig. 12b the same network as in Fig. 12a appears, but with a set of currents and voltages $[\bar{e},\bar{i}]$ that forms a static solution of the dynamic model. (This would also be a solution of the corresponding static model, that is, the dynamic model with inductors shorted and capacitors open.) The diodes to the left and right of the box are chosen as typical of diodes in the open and closed states for the static solution $[\bar{e},\bar{i}]$.

Next, we construct a network that has as its solution $[e(\tau),i(\tau)]$, the difference between the two solutions just described.

$$[e(\tau),i(\tau)] = [\overset{*}{e}(\tau)-\bar{e},\overset{*}{i}(\tau)-\bar{i}].$$

The difference $[e(\tau),i(\tau)]$ automatically satisfies the demands of all components inside the box by virtue of their linearity. The constraints at the voltage or current source terminals of the box are readily met by substituting short circuits or open circuits, respectively, as indicated in Fig. 12c. All resistances have also been brought outside the box for the arguments that will be made below. In the case of the diode ports care must be exercised: First consider a diode that is in the closed state in the static solution, $\bar{e}_d = 0$. For those segments of time in which this diode is closed in the solution of the dynamic model ($\overset{*}{e}_d(\tau)=0$), a short circuit may be substituted for the diode in the difference network. For those segments in which the diode is not closed ($\overset{*}{e}_d(\tau) > 0$), we may substitute for the diode a current source of value $\bar{i}_d$.

A similar argument holds with regard to a diode that is open in the static solution: In the difference network the diode is replaced by an open circuit in periods where $\overset{*}{i}_d = 0$, and by a voltage source $\bar{e}_d$ in periods where $\overset{*}{i}_d > 0$.

Consider the electrical power leaving the box in the difference network. At a resistance port, the power can never be negative, and is greater than zero if either i_r or e_r in nonzero. For a diode port, the power leaving the box is zero except when both i_d and e_d are nonzero. For a closed diode we have

$$i_d = \bar{i}_d \geqslant 0$$

$$e_d = \overset{*}{e}_d \leqslant 0$$

and thus

$$-e_d i_d \geqslant 0 \ .$$

The power leaving the box through a diode port in the difference solution is never negative and is greater than zero if and only if the product $e_d i_d$ is nonzero. Since the total power flow out of the box is never negative, the stored energy in the inductors and capacitors of the difference network must monotonically decrease as time passes. But the stored energy can never become negative, and it follows that a limit point (not necessarily zero) must be approached which is a lower bound on the stored energy in the difference network. As the stored energy approaches its lower bound asymptotically, the power flow out of the network through the diode and resistance ports must approach zero. This statement may be rigorously established by an argument based on the uniform continuity of the solution $[\overset{*}{e}(\tau), i^*(\tau)]$ at all but a countable number of points in time.

Thus the difference solution $[e(\tau), i(\tau)]$ must tend toward a behavior for which $e_d i_d$ is zero for each diode and $e_r i_r$ is zero for each resistance. If the natural frequencies of the network (apart from any lying at the origin) have negative real parts for all combinations of diode states, the limiting behavior must be a static solution of the difference network. For the dynamic model of the network flow problem, these natural frequencies are all dissipative, since each inductor is paralleled with a resistance and each capacitor is in series with a resistance.

Thus any solution of the difference network tends to a static limit, and the dynamic solution $[\overset{*}{e}(\tau), \overset{*}{i}(\tau)]$, must also approach a static solution in the limit. This static solution need not be identical to the solution $[\bar{e}, \bar{i}]$ used in the argument. All static solutions of the dynamic model, however, are equivalent in that they all correspond to equally valid optimal solutions of the network flow problem.

RANGE OF NATURAL FREQUENCIES

The choice of values for the inductors, capacitors and resistors used in the formation of the dynamic model sets the time scale of convergence to the optimal flow pattern. If this time scale is too long in relation to the iteration period δ of the distributed algorithm, excessive iterations will be required. If the time scale is too short in relation to δ, erratic behavior may result. The frequency of iteration necessary is determined principally by the highest frequency component in the solution, while the time for convergence to a usable solution is determined by the lowest frequency component of the transient solution.

Thus it is appropriate to ask what range of natural frequencies (apart from the root at zero frequency) is to be expected in the dynamic model under the various configurations of diode states. Some insight into this problem may be gained by regarding the transient behavior of the dynamic model as produced by a two-element-kind network built up from the parts shown in Fig. 13. If we assume that all components of a given type have the same value, loci of the natural frequencies in the complex frequency plane may be constructed by an elegant analysis.[11] Tellegen's theorem states that the sum of powers delivered to all components of a network is identically zero. In a two-element-kind network, the complex power absorbed by one kind of element must be delivered by elements of the other class, that is, for each natural frequency s of the network

$$Z_a \sum_a |I_a|^2 = - Z_b \sum_b |I_b|^2$$

or

$$-Z_a Y_b = \eta \geqslant 0$$

where η is some non-negative real number. From this condition and the expressions for Z_a and Y_b

$$Z_a(s) = \frac{1}{\gamma} + \frac{1}{\kappa s} \qquad Y_b(s) = \frac{1}{\rho} + \frac{1}{\lambda s}$$

we derive a pair of equations constraining the real and imaginary parts of $s = \sigma + j\omega$. It is assumed that the two impedances have equal time constants.

$$\frac{\lambda}{\rho} = \frac{\kappa}{\gamma} = \theta$$

The equations are

$$(\sigma^2-\omega^2)\ \theta^2(i+\eta\ \rho\ \gamma) + 2\ \sigma\ \theta + 1\ 0 \tag{4}$$

$$\omega[\sigma\ \theta^2(1+\eta\ \rho\ \gamma) + \theta] = 0 \tag{5}$$

From Equation (5) we see that two cases may prevail;

1) $\omega = 0$

for which (4) yields a real value for σ

$$\sigma = - \frac{1}{\theta}$$

only when $\eta = 0$; and

2) $\omega \neq 0$

where we find

$$\sigma = -\frac{1}{\theta(1+\eta\rho\gamma)}$$

and

$$\omega = \pm\frac{\eta\rho\gamma}{\theta(1+\eta\rho\gamma)}$$

where η is real and non-negative. Figure 14 is a plot of these results indicating the circular locus on which the natural frequencies of the dynamic model must lie. A clear upper bound on the range of natural frequencies is established by this analysis

$$-\frac{1}{\theta} \leqslant \sigma \leqslant 0$$

A lower bound would be determined if an upper bound could be placed on η. The maximum value attainable by η is dependent on the number of elements in the network and their configuration, increasing with the complexity of the network. The nature of this dependence is a matter for further study.

THE MULTI-COMMODITY PROBLEM

The multi-commodity flow problem differs from the network flow problem in that flow in the links must be divided into categories which separately satisfy the flow continuity condition, and must be individually non-negative. The problem may be regarded in terms of a layered network in which each of the identical layers accommodates non-negative flow of one category. The constraint on maximum flow in a link now applies to the sum of all categories of flow in the link. If we use the subscript k to index the different categories of flow and make the obvious estensions of the notation used for the network flow problem, the mathematical statement of the multi-commodity flow problem is

minimize

$$\phi = \sum_{i,k} c_i x_{ik} \qquad \text{(total cost)}$$

with

$$\sum_k x_{ik} + u_i = m_i \qquad \text{(flow bounds)}$$

$$x_{ik} \geqslant 0,\ u_i \geqslant 0$$

and

$$\sum_{i \ni O(i)=j} x_{ik} - \sum_{i \ni T(i)=j} x_{ik} = f_{jk} \qquad \text{(continuity)}$$

By the same procedure used before, a Lagrangian statement of the multi-commodity problem can be constructed and related to an electrical model which is shown in Fig. 15. One additional component, the ideal transformer, is necessary to realize this model and is defined in the figure.

The distributed algorithm concept developed in this paper is readily applied to the more general problem and yields a set of computation rules only slightly different in appearance from those given in Fig. 11. The theoretical arguments given earlier regarding convergence and range of natural frequencies are unaffected by the inclusion of transformers in the electrical model. Standard techniques for the network flow problem, on the contrary, are not easily extended to handle the multi-commodity problem. This is only one way in which the distributed solution technique suggested here can be extended. Other immediate extensions are to flow networks with gain and networks with nonlinear costs.

Acknowledgement

The work reported here is a direct extension of ideas developed while the author served as consultant to ADCOM, Inc., in Cambridge, Massachusetts. Appreciation is expressed to Dr. Elie J. Baghdady and other colleagues at ADCOM for introducing the author to the problem of message routing in communication systems and permitting him the freedom to explore his ideas.

In the preparation of this report, the writer has benefited greatly from a number of stimulating discussions with Professor Harry B. Lee whose suggestions form the basis of the convergence argument and the natural frequency analysis.

References

This work was supported in part by the U. S. Army Signal Corps, the Air Force Office of Scientific Research, and the Office of Naval Research; in part by the National Science Foundation (Grant G-16526), the National Institutes of Health (Grant MH-04737-03), and the National Aeronautics and Space Administration (Grant NsG-496).

The work was performed in part at ADCOM, Inc., Cambridge, Mass., under subcontract with ITT Communications Systems, Inc., Paramus, N.J.

[†] Department of Electrical Engineering and Research Laboratory of Electronics, Massachusettes Institute of Technology, Cambridge, Mass.

[‡] We use lower-case Roman letters to represent column vectors and matrices, and Greek letters to denote scalars. Matrix transposition is indicated by superscript T.

[1] L. R. Ford and D. R. Fulkerson, "Flows in Networks," Princeton University Press, Princeton, N.J.: 1962.

[2] F. L. Hitchcock, "The distribution of a product from several sources to numerous locations," J. Mathematics and Phys., Vol. 20, pp. 224-230; 1941.

[3] J. T. Rubacker, "Concerning multicommodity networks," RAND Report RM-1799; September, 1956.

[4] J. B. Dennis, "Message routing in communications networks," Working papers in network control strategies, ADCOM, Inc., Cambridge, Mass.; August, 1962.

[5] L. R. Ford and D. R. Fulkerson, " A simple algorithm for finding maximal network flows and an application to the Hitchcock problem," Canad. J. Math., Vol. 9, pp. 210-218; 1957.

[6] D. R. Fulkerson, "An out-of-kilter method for minimal cost flow problems," J. Soc. Indust. App. Math., vol. 9, pp. 18-27; 1961.

[7] J. B. Dennis, "Mathematical Programming and Electrical Networks," John Wiley and Sons, Inc., New York; 1959, Chapter 4.

[8] R. E. Kalaba and M. L. Juncosa, "Optimal design and utilization of communication networks," Management Science, vol. 3, pp. 33-44; 1956.

[9] J. B. Dennis, "Mathematical Programming and Electrical Networks," _op. cit._, Chapter 3.

[10] A. J. Goldman and A. W. Tucker, "Theory of linear programming," in "Linear Inequalities and Related Systems," H. Kuhn and A. Tucker, Eds., Princeton University Press, Princeton, N.J.; 1956.

[11] R. D. Thornton, "Some Limitations of Linear Amplifiers," Sc.D.Thesis, Department of Electrical Engineering, M.I.T.; 1957.

Figure Captions

Fig. 1 - A linear graph.
Fig. 2 - Notation for the network flow problem.
Fig. 3 - Equivalent formulations of the linear programming problem.
Fig. 4 - Primal, dual, and Lagrangian formulation of the network flow problem.
Fig. 5 - Electrical components used in the model.
Fig. 6 - Electrical model of one link of a flow network.
Fig. 7 - Electrical model of the network flow problem.
Fig. 8 - Modified electrical model.
Fig. 9 - Dynamic model for the network flow problem.
Fig. 10- Relations describing the dynamic model of the network flow problem.
Fig. 11- Computation steps for station g.
Fig. 12- Construction of difference solution for convergence discussion.
Fig. 13- Network parts for natural frequency analysis.
Fig. 14- Locus of natural frequencies for the dynamic model.
Fig. 15- Electrical model for the multi-commodity flow problem.

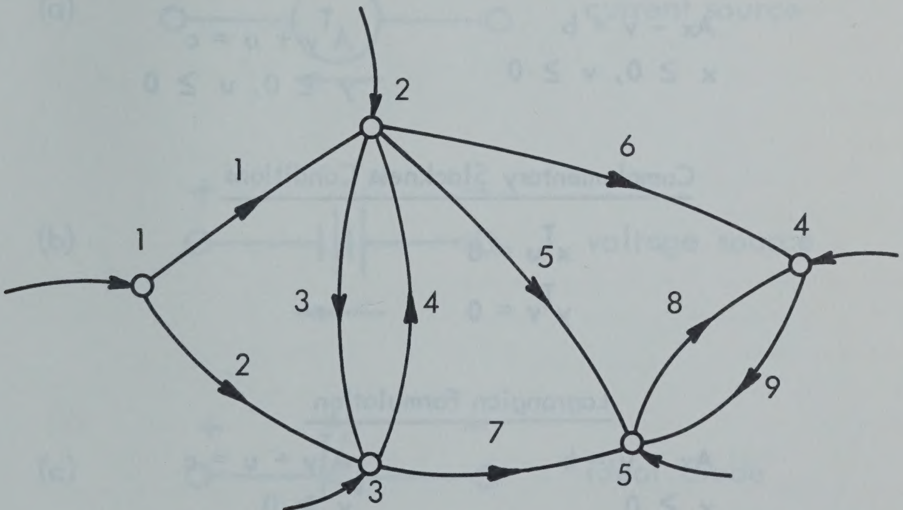
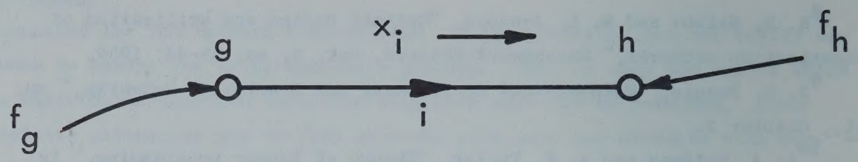


Fig. 1.



$g = O(i)$ origin station

$h = T(i)$ terminal station

x_i flow

c_i unit cost

m_i maximum flow

} in link i

f_g source flow at station g

Fig 2

Primal	**Dual**
Minimize $c^T x$	Maximize $b^T y$
$Ax - v = b$	$A^T y + u = c$
$x \geq 0,\ v \geq 0$	$y \geq 0,\ u \geq 0$

Complementary Slackness Conditions

$$x^T u = 0$$

$$y^T v = 0$$

Lagrangian Formulation

$Ax - v = b$	$A^T y + u = c$
$x \geq 0$	$y \geq 0$
$v \geq 0$	$u \geq 0$

$$x^T u = 0$$

$$y^T v = 0$$

Fig 3

Primal

Minimize

$$\Phi = \sum_i c_i x_i$$

$$x_i + u_i = m_i$$

$$\sum_{O(i)=j} x_i - \sum_{T(i)=j} x_i = f_j$$

$$x_i \geq 0, \; u_i \geq 0$$

Dual

Maximize

$$\theta = \sum_j f_j y_j - \sum_i m_i v_i$$

$$y_g + w_i - v_i - y_h = c_i,$$

$$y = O(i), \; h = T(i)$$

$$w_i \geq 0, \; v_i = 0$$

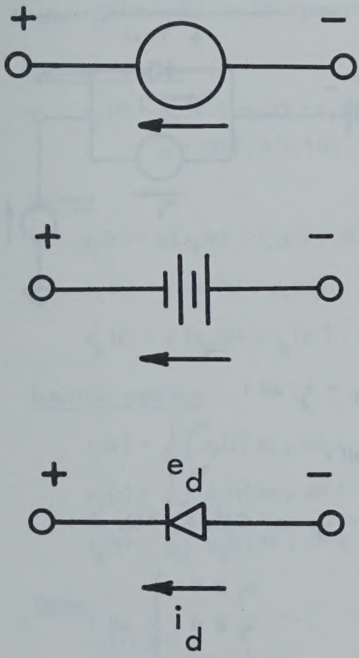
Complementary slackness

$$w_i x_i = 0$$

$$v_i u_i = 0$$

Lagrangian Statement

Fig 4



(a) current source

(b) voltage source

(c) ideal diode

$$e_d \geq 0 \qquad i_d \geq 0$$

$$e_d i_d = 0$$

Fig 5

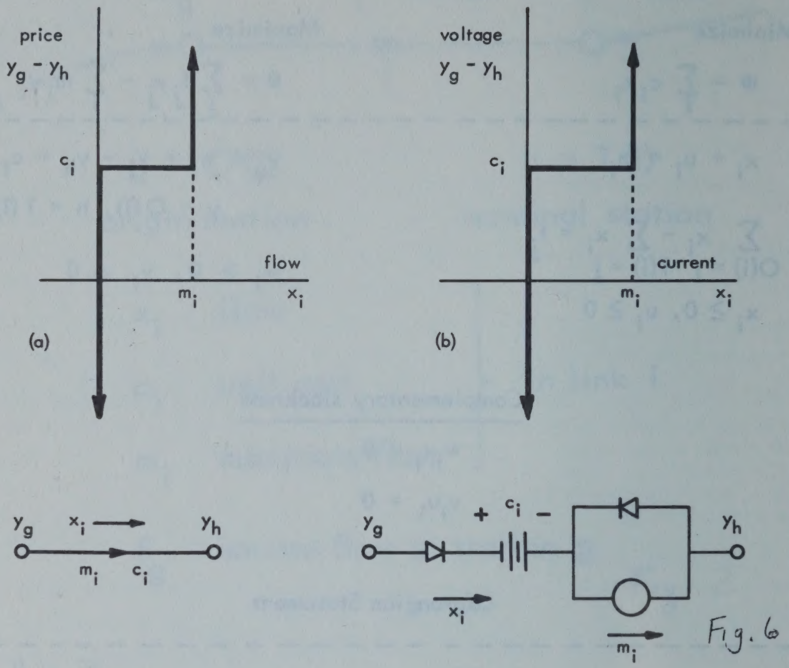
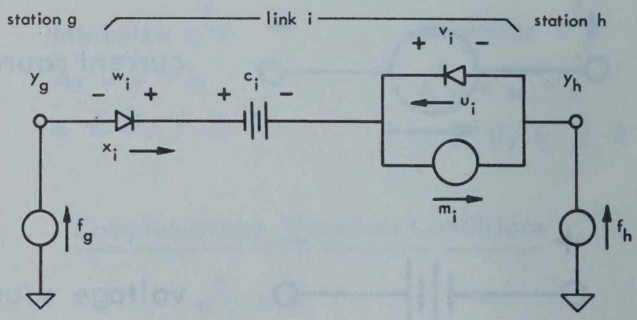

price
$y_g - y_h$
c_i
flow
m_i
x_i
(a)
voltage
$y_g - y_h$
c_i
current
m_i
x_i
(b)
y_g
x_i
y_h
m_i
c_i
y_g
+
c_i
−
y_h
x_i
m_i

Fig. 6

(a)


station g
link i
station h
y_g
w_i
− +
+
c_i
−
x_i
+
v_i
−
u_i
y_h
m_i
f_g
f_h

(b)

Node law:

$$\sum_{O(i)=j} x_i - \sum_{T(i)=j} x_i = f_j, \text{ all } i$$

$$x_i + u_i = m_i, \text{ all } i$$

Loop law:

$$y_g + w_i - v_i - y_h = c_i, \quad g = O(i),\ h = T(i), \text{ all } i$$

Diodes:

$$\left.\begin{array}{l} x_i \geq 0 \\ w_i \geq 0 \\ x_i w_i = 0 \end{array}\right\} \text{ all } i \qquad \left.\begin{array}{l} u_i \geq 0 \\ v_i \geq 0 \\ u_i v_i = 0 \end{array}\right\} \text{ all } i$$

Fig 7.

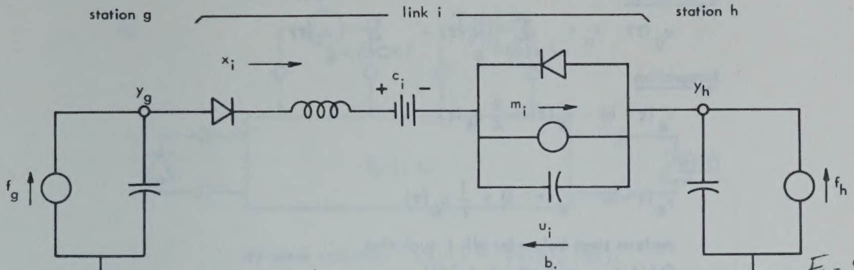

station g
link i
station h
x_i
y_g
c_i
+
−
m_i
y_h
f_g
f_h
u_i

Fig 8

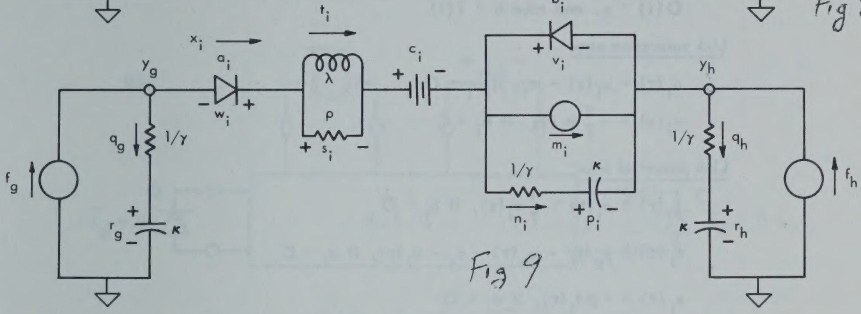

b_i
t_i
x_i
a_i
y_g
c_i
+
−
v_i
y_h
w_i
λ
ρ
s_i
q_g
1/γ
m_i
f_g
1/γ
κ
n_i
p_i
r_g
q_h
r_h
f_h

Fig 9

Kirchhoff node law

$$\sum_{i \ni O(i)=g} x_i(\tau) - \sum_{i \ni T(i)=g} x_i(\tau) + q_g(\tau) = f_g\,, \text{ all } g$$

$$x_i(\tau) + u_i(\tau) - n_i(\tau) = m_i\,, \text{ all } i$$

Kirchhoff loop law

$$y_g(\tau) - y_h(\tau) + w_i(\tau) - v_i(\tau) - s_i(\tau) = c_i\,,$$
$$g = O(i)\,,\ h = T(i)\,,\quad \text{all } i$$

Resistors

$$s_i(\tau) = \rho\,[x_i(\tau) - t_i(\tau)]\ , \text{ all } i$$

$$n_i(\tau) = \gamma\,[v_i(\tau) - p_i(\tau)]\ , \text{ all } i$$

$$q_g(\tau) = \gamma\,[y_g(\tau) - r_g(\tau)]\ , \text{ all } g$$

Inductors, capicitors

$$t_i(\tau) = \frac{1}{\lambda}\int^{\tau} s_i(\tau)\,d\tau\ , \text{ all } i$$

$$p_i(\tau) = \frac{1}{\kappa}\int^{\tau} n_i(\tau)\,d\tau\ , \text{ all } i$$

$$r_g(\tau) = \frac{1}{\kappa}\int^{\tau} q_g(\tau)\,d\tau\ , \text{ all } g$$

Diodes

$$\left.\begin{aligned} x_i(\tau) &\geq 0 \\ w_i(\tau) &\geq 0 \\ x_i(\tau)\,w_i(\tau) &= 0 \end{aligned}\right\} \text{all } i \qquad \left.\begin{aligned} u_i(\tau) &\geq 0 \\ v_i(\tau) &\geq 0 \\ u_i(\tau)\,v_i(\tau) &= 0 \end{aligned}\right\} \text{all } i$$

Fig 10.

Error in flow

$$q_g(\tau) = f_g + \sum_{i \ni T(i) = g} x_i(\tau) - \sum_{i \ni O(i) = g} x_i(\tau)$$

Integration

$$r_g(\tau + \delta) = r_g(\tau) + \frac{\delta}{\kappa} q_g(\tau)$$

Price

$$y_g(\tau + \delta) = r_g(\tau + \delta) + \frac{1}{\gamma} q_g(\tau)$$

perform steps below for all i such that
$O(i) = g$, and take $h = T(i)$

Link saturation error

$$n_i(\tau) = x_i(\tau) - m_i, \text{ if } b_i = O$$

$$n_i(\tau) = -\frac{1}{\gamma} p_i(\tau), \text{ if } b_i = C$$

Link potential error

$$v_i(\tau) = p_i(\tau) + \frac{1}{\gamma} n_i(\tau), \text{ if } b_i = O$$

$$s_i(\tau) = y_g(\tau) - y_h(\tau) - c_i - v_i(\tau), \text{ if } a_i = C$$

$$s_i(\tau) = -\rho t_i(\tau), \text{ if } a_i = O$$

Integration

$$t_i(\tau + \delta) = t_i(\tau) + \frac{\delta}{\lambda} s_i(\tau)$$

$$p_i(\tau + \delta) = p_i(\tau) + \frac{\delta}{\kappa} n_i(\tau)$$

Flow

$$x_i(\tau + \delta) = t_i(\tau + \delta) + \frac{1}{\rho} s_i(\tau), \text{ if } a_i = 1$$

Dependent quantities

$$w_i(\tau + \delta) = -\left[y_g(\tau + \delta) - y_h(\tau + \delta) - c_i \right], \text{ if } a_i = O$$

$$v_i(\tau + \delta) = y_g(\tau + \delta) - y_h(\tau + \delta) - c_i - \rho t_i(\tau + \delta), \text{ if } b_i = O$$

$$u_i(\tau + \delta) = m_i + n_i(\tau + \delta) - x_i(\tau + \delta), \text{ if } a_i = C \text{ and } b_i = C$$

State of link

if $a_i = O$ and $w_i(\tau + \delta) < 0$;

$$w_i(\tau + \delta) = 0, \quad C \rightarrow a_i$$

if $b_i = O$ and $v_i(\tau + \delta) < 0$;

$$v_i(\tau + \delta) = 0, \quad C \rightarrow b_i$$

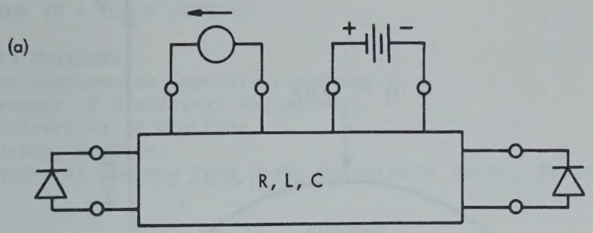
if $a_i = C$, $b_i = C$ and $x_i(\tau + \delta) < 0$;

$$x_i(\tau + \delta) = 0, \quad O \rightarrow a_i$$

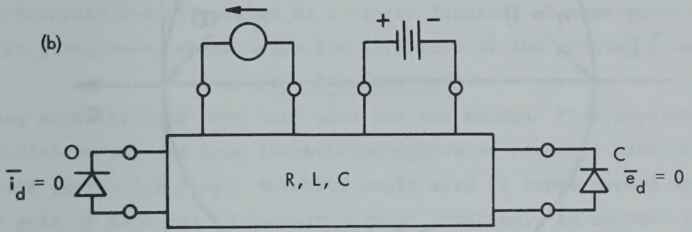
if $a_i = C$, $b_i = C$ and $u_i(\tau + \delta) < 0$;

$$u_i(\tau + \delta) = 0, \quad O \rightarrow b_i$$

Fig 11



dynamic solution: $[e, i] = [\overset{*}{e}(\tau), \overset{*}{i}(\tau)]$


(b)
+
−
$\bar{i}_d = 0$
R, L, C
C
$\bar{e}_d = 0$

static solution: $[e, i] = [\bar{e}, \bar{i}]$

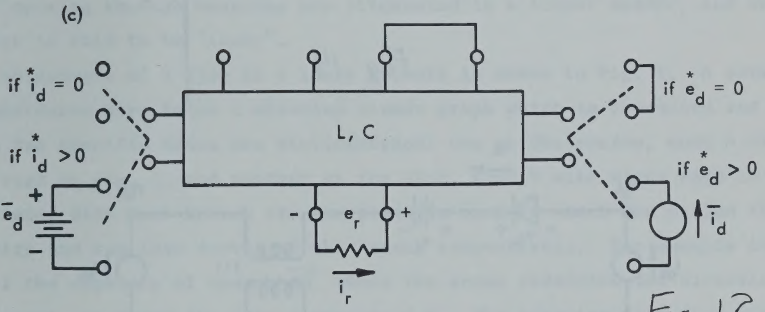

(c)
if $\overset{*}{i}_d = 0$
if $\overset{*}{i}_d > 0$
+
$\bar{e}_d$
−
L, C
−
e_r
+
i_r
if $\overset{*}{e}_d = 0$
if $\overset{*}{e}_d > 0$
$\bar{i}_d$

Fig. 12

difference solution: $[e, i] = [\overset{*}{e}(\tau) - \bar{e}, \overset{*}{i}(\tau) - \bar{i}]$

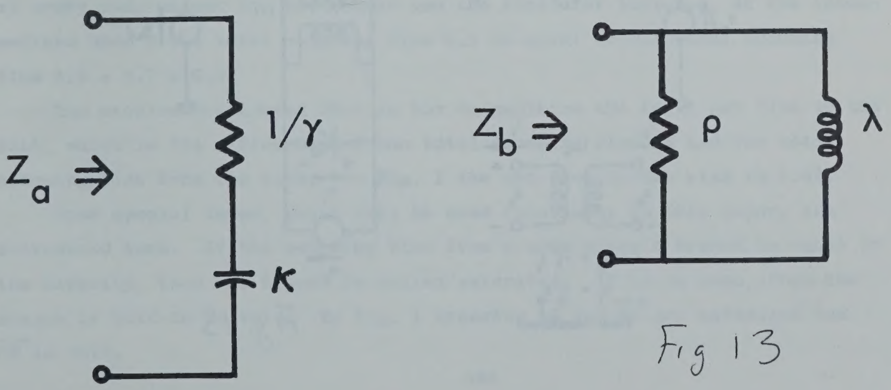

$Z_a \Rightarrow$
$1/\gamma$
K
$Z_b \Rightarrow$
ρ
λ

Fig 13

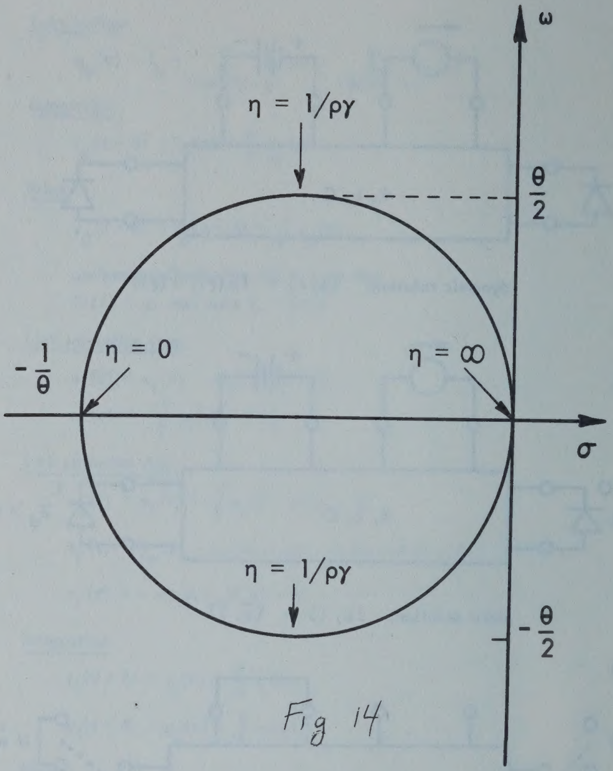

ω
η = 1/ργ
$\frac{\theta}{2}$
$-\frac{1}{\theta}$
η = 0
η = ∞
σ
η = 1/ργ
$-\frac{\theta}{2}$

Fig 14

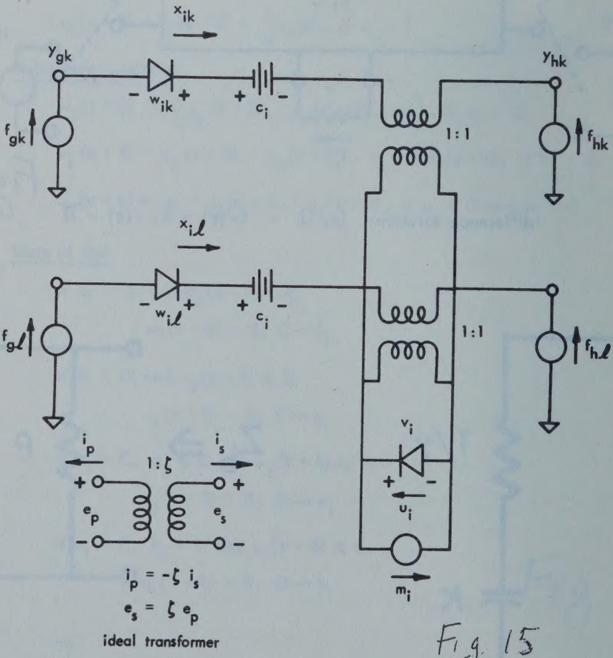

x_{ik}
y_{gk}
y_{hk}
w_{ik}
c_i
f_{gk}
f_{hk}
1:1
$x_{i\ell}$
$w_{i\ell}$
c_i
$f_{g\ell}$
$f_{h\ell}$
1:1
v_i
u_i
m_i
i_p
i_s
1:ζ
e_p
e_s
$i_p = -\zeta\, i_s$
$e_s = \zeta\, e_p$
ideal transformer

Fig 15

MAXIMAL FLOW IN A LOSSY NETWORK

TOSHIO FUJISAWA
Cooley Electronics Laboratory and the
Department of Electrical Engineering
The University of Michigan
Ann Arbor, Michigan
(On leave of absence from Osaka University, Osaka, Japan)

Summary

This paper is concerned with the problem of maximizing the flow in a network with a source and a sink, which is lossy in the sense that flows passing through branches are attenuated in a linear fashion. A topological solution method is given here, which makes the structure of the problem clear.

I. Introduction

Many contributions have been made for the maximal flow problem in networks, which we may see from the bibliographies of recently published excellent books [Refs. 1,2,3,4]. Networks dealt with in these papers assume no loss or gain of flow due to passage through branches. In certain networks such as those for power or information transmission, it is more realistic to introduce loss factors. In this paper loss factors are introduced so that flows passing through branches are attenuated in a linear manner, and such a network is said to be "lossy".

An example of a flow in a lossy network is shown in Fig. 1. A network is considered here to be a directed linear graph which is connected and in which two specific nodes are distinguished; one as the source, node A with plus sign in Fig. 1, and another as the sink, node D with minus sign in Fig.1. Associated with each branch are two positive numbers which are called the capacity and the loss factor of the branch respectively. For example in Fig. 1 the capacity of branch $\overrightarrow{AB}$, where the arrow indicates the direction of the branch, is 5 and the loss factor is 0.9. The outgoing flow from node A in branch $\overrightarrow{AB}$ cannot exceed the capacity 5 and is actually 5 in Fig. 1 and the incoming flow to node B is 4.5 which is the loss factor multiple of the outgoing flow from node A. We should notice that the conservation law must hold at every node except for the source and the sink; for instance, at the intermediate node C the total outgoing flow 6.3 is equal to the total incoming flow 3.6 + 2.7 = 6.3.

The problem considered here is how to maximize the total net flow to the sink, which is the difference of the total incoming flow to and the total outgoing flow from the sink. In Fig. 1 the net flow to the sink is 5.67.

Some special terms, which will be used constantly in this paper, are introduced here: If the outgoing flow from a node along a branch is equal to the capacity, then the branch is called saturated. If it is zero, then the branch is said to be void. In Fig. 1 branches $\overrightarrow{AB}$ and $\overrightarrow{AC}$ are saturated and $\overrightarrow{BD}$ is void.

II. Lossy Networks

There may be many paths connecting between the source and the sink in a network. When we try to increase flows from the source along these paths, sooner or later we reach a situation in which no more increase of flows is possible because all the branches are capacitated. In such a situation there appears a cut set separating between the source and the sink as shown in Fig. 2, in which all the forward-directed branches of the cut set are saturated and all the backward-directed branches of the cut set are void. The cut set splits all the nodes up to two groups S^+ and S^- as shown in Fig. 2, in which S^+ contains the source but not the sink. By forward-directed branches we mean branches directed from nodes of S^+ to nodes of S^-, and by backward-directed branches we mean reversely directed branches.

Now we can say S^- is receiving the maximum possible supply from the outside. No other flows can give any more supply to S^-. In the usual maximal flow problem without loss factors such a flow is really maximal; because there is no loss, the total supply from the outside to S^- is just equal to the netsupply to the sink. However, this is not the case in our lossy networks.

We will denote by W the subgraph consisting of all the nodes of S^- and all the branches connecting between nodes of S^-, and of course assume W is connected. Since W is connected, there exists at least one (complete) tree T^- of W [Refs. 2,5]. It is easy to see that flows in T^- are uniquely determined by the flow conservation law if flows in all the branches, that belong to W but not to T^-, are fixed.

Let us now give certain numbers, referred to as potentials, to all the nodes of T^- as follows: Define the potential of the sink as unity. If either one of the potentials of the terminating nodes A and B of a branch $\overrightarrow{AB}$ of T^- has been already defined, then define the remaining one by the following relation (see Fig. 3)

$$\mu_A = \alpha \mu_B, \tag{1}$$

where μ_A and μ_B stand for the potentials of nodes A and B respectively, and where α stands for the loss factor of the branch. There is only one path connecting between the sink and any node of T^- for T^- is a tree [Ref. 5]. Therefore, by induction the above definition uniquely determines all the potentials.

Let us consider what potentials indicate. An example of potentials is given in Fig. 4. Let us suppose that node A is given an excess supply of quantity Δ from outside. Then, to keep the conservation law at node A the outgoing flow from node B along the branch $\overrightarrow{BA}$ is decreased by the amount $\Delta/0.6 = \Delta \mu_A/\mu_B$ and this change gives rise to an excess supply of the same quantity at node B. Again this excess supply at node B is compensated by increasing the flow along the branch $\overrightarrow{BC}$ by the same amount $\Delta \mu_A/\mu_B$, and this change causes the increase of the net flow to the sink C by the amount $(\Delta\, \mu_A/\mu_B)\mu_B = \mu_A\Delta$. Now the meaning of potentials has been made clear. An excessof supply Δ at node A gives rise to the increase of the net flow to the sink by the amount $\Delta\, \mu_A$ where μ_A stands for the potential of node A.

The idea of potentials provides us an efficient means to judge what flows in W are most efficient and maximality criteria are expressed in terms of potentials in the following section.

III. Maximality Criteria

Following the discussions in the preceding section, let us consider a directed branch $\overrightarrow{AB}$ that belongs to W but not to T^-. There is only one path, in T^-, connecting between node A and the sink. The same is true with node B. Therefore, these two paths together with branch AB constitute a subgraph such as is shown in Fig. 5.

Let us now see the effect of increasing the flow in branch $\overrightarrow{AB}$ by a quantity X. This means that there is a shortage of supply X at node A and the excess of supply αX at node B. According to the discussion on potentials the former results in a decrease of supply $X\mu_A$ to the sink and the latter results in an increase of supply $\alpha X\mu_B$ to the sink. Therefore, the net effect is a change of supply

$$(\alpha\, \mu_B - \mu_A)\, X \tag{2}$$

to the sink.

If $\overrightarrow{AB}$ is a saturated branch, then the increase of the net flow to the sink is made possible only when $\alpha\, \mu_B - \mu_A < 0$ because X can be only negative. If $\overrightarrow{AB}$ is void, then the increase of the net flow to the sink is possible only when $\alpha\, \mu_B - \mu_A > 0$. If $\overrightarrow{AB}$ is neither saturated nor void, values of X can be positive and also negative and hence an increase of the net flow to the sink is always possible except for the special case $\alpha\, \mu_B - \mu_A = 0$.

If the quantity (2) cannot be made positive for every branch that belongs to W but not to T^-, then there is no way to increase the net flow to the sink by changing flows in W. As we noticed earlier, W is already receiving the maximum possible supply from the outside. Therefore, the current net flow to the sink must be maximal. In other words the flow is maximal if the following conditions are met: 1) If $\alpha\, \mu_B > \mu_A$ for a branch $\overrightarrow{AB}$ in W, the branch is saturated, and 2) if $\alpha\, \mu_B < \mu_A$ for a branch AB in W, the branch is void. The maximality conditions given here are sufficient but not necessary.

The conditions imply that only flows such that all the branches in W but not in T^- are saturated or void are worthwhile to consider for finding a maximal flow. Fulkerson and Dantzig proposed in their joint paper [Ref. 6] to confine themselves to such flows that 2-trees [Ref. 2] separating the source from the sink can be formed after discarding all the saturated branches and some of the void branches. However, the 2-tree concept is not sufficient in our lossy case. Markowitz [Ref. 7]showed that certain kinds of loops are very important to the lossy case.

Let us suppose we are increasing the value of χ from zero for the case $\alpha\mu_B - \mu_A > 0$. Because of the existence of branch capacities, it is impossible to increase the value of χ indefinitely. There should be an upper limit to the value of χ, and when the value of χ reaches this limit at least one branch in Fig. 5 becomes saturated or void. If the branch is, for example, EF, then there remains a loop after discarding the branch. This loop, however, has a remarkable property if all of the branches of the loop are neither void nor saturated. Such a loop can behave as a source. A shortage of supply χ at node A results in a decrease of supply $\chi \cdot \mu_A$ at the sink, which is equivalent to saying that there results a shortage of supply $\chi\mu_A/\mu_E$ at node E in Fig. 5 as long as a compensation along the path from node E to the sink is not made. Also, in the same sense, the excess of supply $\alpha\chi$ at node B results in an excess of supply $\alpha \cdot \chi\mu_B/\mu_E$ at node E in Fig. 5 as long as compensation is not made along the path from E through F to the sink. Therefore, at node E there appears a net excess of supply

$$(\alpha\mu_B - \mu_A)\chi/\mu_E \tag{3}$$

As long as $\alpha\mu_B \neq \mu_A$ and the value of χ can be made positive and negative without violating capacity constraints, the quantity (3) can be made positive and the loop looks like a source.

Now let us consider a loop $A_1A_2\ldots A_nA_{n+1}$ ($A_{n+1} = A_1$) as shown in Fig. 6. If we discard one branch, say A_iA_{i+1}, then there remains a tree. Therefore, we can define potentials $\mu_{A_1}, \ldots, \mu_{A_n}$ starting with the definition, for example $\mu_{A_1} = 1$, just as in the preceding section. Let us denote the loss factor of branch A_iA_{i+1} by α. We will call the loop regular if branch A_iA_{i+1} is directed as $\overrightarrow{A_iA_{i+1}}$ and $\mu_{A_i} \neq \alpha\mu_{A_{i+1}}$ or if branch A_iA_{i+1} is directed reversely and $\mu_{A_{i+1}} \neq \alpha\mu_{A_i}$.

It is easily seen that whether a loop is regular or not does not depend on the choice of the discarded branch or of the starting point at which the potential is set at unity. Any node of a regular loop can behave as a source by changing the flows in the loop as long as all the branches of the loop are neither saturated nor void.

We are now ready to introduce a solution method which depends on the tree concept as well as the regularity concept of loops.

IV. Solution Method

Suppose we are given a flow in a lossy network. Let us denote by G a subgraph obtained by discarding some of the saturated branches and some of the voide branches. The flow is called basic with geometry G if there exists a subgraph G satisfying the following conditions: 1) The connected component [Ref. 5] T^+ containing the source is a tree and does not contain the sink, 2) the connected component T^- containing the sink is a tree, and 3) each of other components L_1, L_2, ..., if they are present, contains one and only one loop that is regular.

A subgraph G satisfying the above conditions is referred to as a geometry of the basic flow. For a given flow there may be more than one geometry and hence more than one basic flow. It is easy to see that the assignment of a geometry G and the set of saturated branches, that do not belong to G, uniquely determines flows in all branches, and therefore there are only a finite number of basic flows.

The solution method to be given here is an iterative procedure which starts with an initial basic flow and produces a new basic flow without decreasing the net flow to the sink if the maximality criteria are not met. If the infinite cycling among basic flows does not occur, the iterative procedure always arrives at a basic flow, which satisfies the maximality criteria, after a finite number of iterations because the number of basic flows is finite.

It is very simple to get an initial basic flow with which to start the iteration. Since the network is connected, there is a 2-tree of the network [Ref. 2]. No flows at all with the 2-tree as the geometry furnishes the desired initial basic flow.

Now we can discuss the production of a new basic flow when a given basic flow does not satisfy the maximality criteria. First consider the case in which T^- is not receiving the maximum possible supply from the outside. Suppose there is a branch $\overrightarrow{AB}$ which is directed from node A of T^+ to node B of T^- and which is not saturated. Then, the path, in T^+, connecting between the source and node A and the path, in T^-, connecting between the sink and node B together with branch AB constitute a path connecting between the source and the sink as shown in Fig. 7. We are going to send out a flow from the source along the path. Since all the branches have capacities there is a limit to the possible increase of the net flow to the sink. When this limit is reached, some of the branches of the path become saturated or void and only one of them has to be discarded.[1] Thus geometry G is first supplemented by branch AB and then one branch of the path is taken off; the modified G provides the geometry of the new basic flow.[2]

Secondly let us suppose there is a branch $\overrightarrow{AB}$ which is directed from node A of L_i to node B of T^- and which is not saturated. Then, the regular loop in L_i and the path, in L_i, connecting between the loop and node A and the path, in T^-, connecting between the sink and node B together with branch AB constitute a subgraph as shown in Fig. 8.

In Fig. 8 node E, that is the junction point of the loop and the path, behaves like a source. There is also a limit for the value which can be sent out from the junction E. When the value assumes the limit, some of the branches in Fig. 8 become saturated or voide and only one of them has to be discarded.[1] Thus geometry G is first modified by adding a branch AB and then removing a branch of the loop or the path. The modified G provides the geometry of the basic flow.[2]

---

[1] If the limit is zero, always discard a branch other than branch $\overrightarrow{AB}$. This can happen if a branch of the path is void and backward-directed, or if a branch of the path is saturated and forward-directed.

The case in which there is a non-void branch $\overrightarrow{BA}$ directed from node B of T^- to node A of T^+, L_1, L_2, ..., may be treated in the same way as mentioned above.

Now let us consider the case in which T^- is receiving the maximum possible supply from the outside but not satisfying the maximality criteria. Then, there is a void branch $\overrightarrow{AB}$, connecting between nodes A and B of T^-, such that $\alpha\, \mu_B > \mu_A$ or there is a saturated branch $\overrightarrow{AB}$, connecting between nodes A and B of T^-, such that $\alpha\, \mu_B < \mu_A$ (see Fig. 5). Obviously this case is just the same as the case treated in Fig. 8, and hence no more mention is made here.

V. Discussions and an Example

First of all we should remark here that loss factors do not mean real losses. Loss factors may exceed unity and may mean gains instead of losses.

In the usual maximal flow problem all loss factors are assumed unity and no regular loops exist. The solution method given here coincides with the method due to Fulkerson and Dantzig [Ref. 6] when all loss factors reduce to unity.

We noticed earlier that the iterative procedure successfully terminates in a maximal flow as long as the limitless cycling among some basic flows does not happen. What theoretically assures the absence of limitless cycling? One of the ways to prevent the cycling is to modify the problem so that we are able to always increase the net flow to the sink when producing a new basic flow from a given basic flow. Let us suppose that there are n branches with capacities c_1, c_2, ..., c_n. The capacity constraint for the i-th branch $\overrightarrow{AB}$ has been (see Fig. 9)

$$0 \leqq X \leqq c_i \,. \tag{4}$$

Now let us modify the constraint as follows:

$$-\epsilon^i \leqq X \leqq c_i + \epsilon^i \,, \tag{5}$$

in which ϵ is a sufficiently small positive number. Also let us consider any flow to be of the form

$$X = X_o + X_1 \epsilon + \dots X_n \epsilon^n \,, \tag{6}$$

that is a polynomial of ϵ and of degree n. If the quantity (6) satisfies the inequality (5) for sufficiently small ϵ, then the constant term X_o satisfies the original constraint (4). Therefore, the only difference between the original and modified problems is in the terms including ϵ. In the modified problem the strict increase of the new flow to the sink is made possible after completing one cycle of the iterative procedure and also the branch to be discarded before completing one cycle is uniquely determined. The maximality criteria and the iterative procedure have to be slightly modified for the modified flow problem. This will be discussed in detail elsewhere [Ref. 8]. Dropping all the terms including ϵ from a maximal flow for the modified problem yields a maximal flow for the original problem.

---

[2] There are three possibilities: 1) No increase of flow with a different geometry, 2) increase of flow with the same geometry, and 3) increase of flow with a different geometry.

The problem dealt with in this paper is only the first half of the whole problem. The second half will be how to minimize the net flow from the source while keeping the net flow to the sink maximum. In other words, the second half is to find the most efficient maximal flow. The second half will be treated elsewhere [Ref. 8].

The problem in this paper may be formulated as a linear programming problem and solved by the application of usual linear programming techniques [Ref. 9]. However, the application of topology provides better understanding and algorithms of the problem just as in the usual maximal flow problem.

At the end of this paper an illustrating example is shown. A computer program to carry out the iterative procedure was written in the language of MAD (Michigan Algorithm Decoder) and tested by IBM 7090 at The University of Michigan Computing Center. The lossy network of the example is shown in Fig. 10(a) in which bold-face branches indicate the branches that belong to the geometry of the initial basic flow. After nine iterations a maximal flow was obtained and is shown in Fig. 10(b) in which bold-face branches indicate the branches of geometry G and branches with crossing bars are saturated. This solution took about ten seconds excluding the compilation time of about fifty seconds.

Acknowledgement

The author expresses his sincere gratitude to Professors Alan B. Macnee and George J. Minty, The University of Michigan, for their valuable criticisms. Also thanks go to people who gave the author comments in an informal talk at The University of Michigan where an earlier version of this paper was presented.

References

1. Ford, L. R., Jr., and Fulkerson, D. R., Flows in Networks, Princeton University Press, Princeton, New Jersey, 1962.
2. Seshu, S., and Reed, M. B., Linear Graphs and Electrical Networks, Addison-Wesley, Reading, Massachusetts, 1961.
3. Berge, C., Théorie des Graphes et ses Applications, Dunod, Paris, France, 1958.
4. Kim, W. H., and Chien, R. T., Topological Analysis and Synthesis of Communication Networks, Columbia University Press, New York, N. Y.,1962.
5. König, D., Theorie der Endlichen and Unendlichen Graphen, Chelsea, New York, N. Y., 1950.
6. Fulkerson, D. R., and Dantzig, G. B., "Computation of Maximal Flows in Networks," Naval Research Logistics Quarterly, Vol. 2, pp. 277-283; 1955.

7. Markowitz, H., "Concepts and Computing Procedures for Certain X_{ij} Programming Problems," Proceedings of the 2nd Symposium in Linear Programming, pp. 509-566; 1955.

8. Fujisawa, T., "Maximal Flows in Generalized Networks," (To be published).

9. Gass, P. I., Theory of Linear Programming, Prentice-Hall, Englewood Cliffs, New Jersey, 1959.

List of Illustrations

Figure 1 Example of lossy flow.
" 2 Cut set.
" 3 Definition of potentials.
" 4 Example of potentials.
" 5 Change of flows.
" 6 Regular loop.
" 7 Path between the source and the sink.
" 8 Path between the sink and the regular loop.
" 9 Capacity constraint.
" 10 Illustrative example.

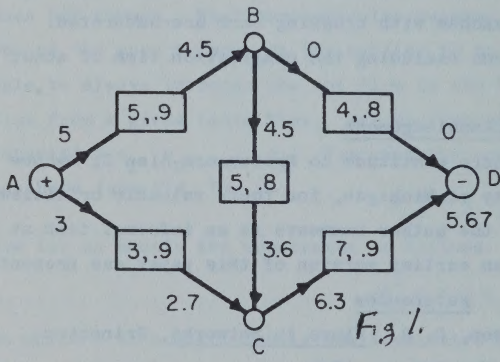

B
4.5
0
5,.9
4,.8
5
4.5
0
A
+
5,.8
−
D
3
5.67
3,.9
3.6
7,.9
2.7
6.3
C

Fig 1.

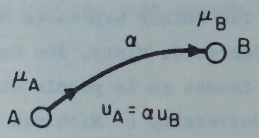

μ_B
α
B
μ_A
A
$u_A = \alpha u_B$

Fig. 3

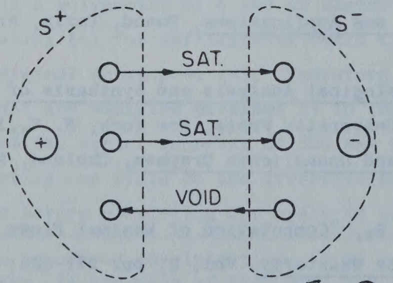

S^+
S^-
SAT.
SAT.
+
−
VOID

Fig. 2

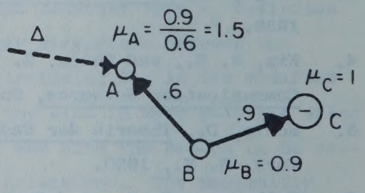

Δ
$\mu_A = \frac{0.9}{0.6} = 1.5$
$\mu_C = 1$
A
.6
.9
−
C
B
$\mu_B = 0.9$

Fig. 4

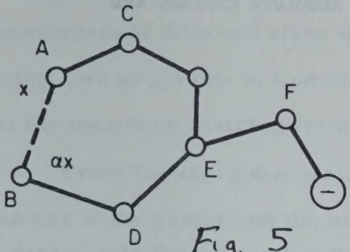


Fig. 5

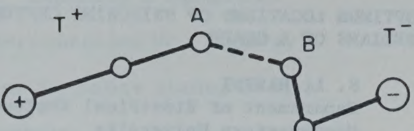


Fig. 7

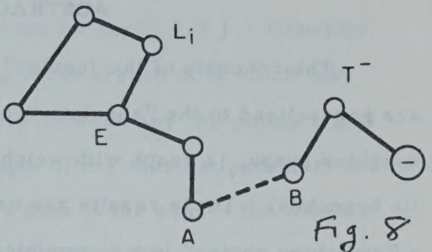


Fig. 8

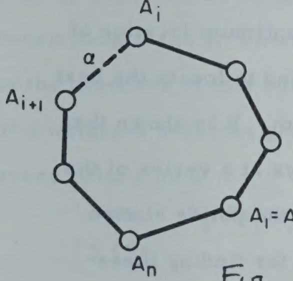
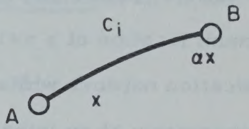
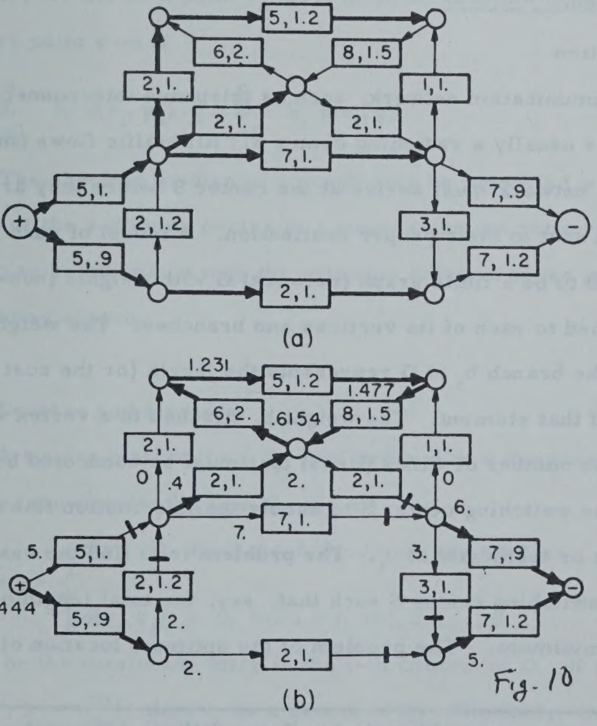

A_i
α
A_{i+1}
A_2
A_1 = A_{n+1}
A_n

Fig. 6



$0 \leq x \leq C_i$

Fig. 9


5, 1.2
6, 2.
8, 1.5
2, 1.
1, 1.
2, 1.
2, 1.
7, 1.
5, 1.
7, .9
+
2, 1.2
3, 1.
−
5, .9
7, 1.2
2, 1.
(a)
1.231
1.477
.6154
0
.4
2.
0
6
7.
5.
3.
4.444
2.
2.
5.
(b)
Fig. 10

MAX NET FLOW TO THE SINK = 11.4

OPTIMUM LOCATIONS OF SWITCHING CENTERS AND THE ABSOLUTE CENTERS AND MEDIANS OF A GRAPH

S. L. HAKIMI
Department of Electrical Engineering
Northwestern University
Evanston, Illinois

ABSTRACT

The concepts of the "center" and the "median vertex" of a graph are generalized to the "absolute center" and the "absolute median" of a weighted graph, (a graph with weights attached to its vertices as well as its branches). These results are used to find the optimum location of a "switching center" in a communication network and to locate the best place to build a "police station" in a highway system. It is shown that the optimum location of a switching center is always at a vertex of the communication network while the best location for the police station is not necessairly at an intersection. Procedures for finding these locations are given.

I. Introduction

In a communication network, such as telephone interconnection system, there is usually a switching center S. All traffic flows (messages) within the network must arrive at the center S before they are processed and then sent to their proper destination. A model of such a system may be assumed to be a finite graph (network) G with weights (non-negative numbers) attached to each of its vertices and branches. The weight w_i attached to the branch b_i of G represents the length (or the cost per unit capacity) of that element. The weight h_i attached to a vertex v_i of G represents the number of wires (lines) that must be connected between vertex v_i and the switching center S to handle the information flows that either originate or terminate at v_i. The problem is to find the exact location of the switching center S such that, say, the total length of the wires is a minimum. The problem of the optimum location of a police

---

*This work was supported by the Air Force Office of Scientific Research, Grant No. AF-AFOSR-98-63 and the U.S. Army Research Office (Durham).

station (or a hospital) is somewhat similar. Consider a number of communities of different sizes that are interconnected by a highway system, we would like to find the location of the police station P such that the maximum distance from P is a minimum.

From the above discussion it is clear that the usual graph theoretic concepts of the center and the median cannot be used [1, 2]. Consider a graph G, a point x on G is a point along some branch of G which may or may not be a vertex of G. The distance between any two points x and y on G, represented by d(x, y), is the length of the shortest path in G between points x and y, where the length of a path is the sum of the weights of the branches of that path [3, 4]. We define a point x_o on an element of a weighted n-vertex graph G to be an absolute center of G, if for every point x on G

$$\max_{1 \leq i \leq n} h_i \, d(v_i, x_o) \leq \max_{1 \leq i \leq n} h_i \, d(v_i, x) \tag{1}$$

Similarly, we define a point y_o on G to be an absolute median of G, if for every point y on G

$$\sum_{i=1}^{n} h_i \, d(v_i, y_o) \leq \sum_{i=1}^{n} h_i \, d(v_i, y) \tag{2}$$

The absolute median of a graph may be identified with the optimum location of the switching center in a communication network, and the absolute center may be identified with the optimum location of the police station in a highway system.

II. Centers and Medians

A square matrix $D = [d_{ij}]$ of order n is a distance matrix of an n-vertex (non-oriented) graph G, if [4]

$$d_{ij} = \begin{cases} d(v_i, v_j), & \text{for } i, j = 1, 2, \ldots, n, \text{ and } i \neq j \\ d(v_i, v_i) = 0, & \text{for } i = j, \; i = 1, 2, \ldots, n \end{cases} \tag{3}$$

Let d_i^m be the maximum entry in the i-th column of D. If $d_c^m = \max (d_1^m, d_2^m, \ldots d_n^m)$, then v_c is a center of G. Similarly, the median

G may be defined, in terms of D, as follows: Let d_i^s be the sum of the entries of the i-th column of G, then if $d_b^s = \max(d_1^s, d_2^s, \ldots, d_n^s)$, v_b is the median vertex of G.

Consider the graph of Fig. 1(a). The distance matrix of that graph is

$$D = \begin{bmatrix} 0 & 10 & 24 & 20 & 34 \\ 10 & 0 & 14 & 12 & 24 \\ 24 & 14 & 0 & 12 & 10 \\ 20 & 12 & 12 & 0 & 20 \\ 34 & 24 & 10 & 20 & 0 \end{bmatrix} \tag{4}$$

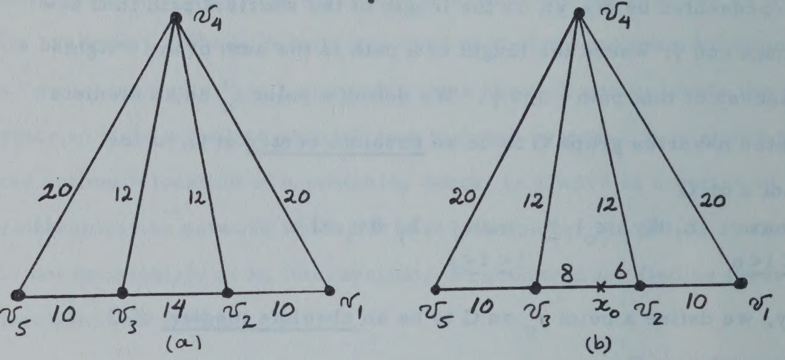


Fig. 1-(a) Center and median of a graph, and (b) the absolute center.

It is easily determined that v_4 is the center of the graph of Fig. 1(a) and that v_2 and v_3 are two possible choices for the median of that graph. The distance of the vertex farthest away from the center of a graph is called the radius r_o of G. In general, we may write

$$r_o = \min_{1 \leq j \leq n} \max_{1 \leq i \leq n} d(v_i, v_j) \tag{5}$$

For example, the radius of the graph of Fig. 1(a) is 20. Let us pick a point x_o on the branch $b(v_2, v_3)$ such that $d(v_2, x_o) = 6$, as shown in Fig. 1(b). It can be seen that for the graph of Fig. 1(b)

$$\max_{1 \leq i \leq 5} d(v_i, x_o) = 18 < d(v_4, v_1) = 20 \tag{6}$$

From the above discussion it follows that there are points on a graph, such as x_o in the graph of Fig. 1(b), which, roughly speaking, are more centerally located than the center of G. The concept of the absolute median is a little more general; because, we have also attached different weights to the vertices of the graph. Similarly, the concept of absolute median is a generalization of the concept of median. The rest of this paper is devoted to describing methods for finding the absolute center and median of a graph.

III. The Absolute Center and the Radius of a Graph

The absolute center of a graph may be redefined as a point x_o on G such that

$$\min_{x \text{ on } G} \max_{1 \leq i \leq n} h_i\, d(v_i, x) = \max_{1 \leq i \leq n} h_i\, d(v_i, x_o) = r(x_o) \tag{7}$$

where $r(x_o)$ is the absolute radius of G. Let the number of branches in G be m, then Eq. (7) may be written, in a more expanded form, as

$$\min_{1 \leq j \leq m} \left[\min_{x \text{ on } b_j} \max_{1 \leq i \leq n} h_i\, d(v_i, x) \right] = r(x_o) \tag{8}$$

However from Eq. (8), we can conclude that the problem of finding the absolute center x_o may be reduced to the problem of solving m simpler "min-max" problems. More precisely, let x_j be a point on b_j such that

$$\min_{x \text{ on } b_j} \max_{1 \leq i \leq n} h_i\, d(v_i, x) = \max_{1 \leq i \leq n} h_i\, d(v_i, x_j) = r(x_j) \tag{9}$$

Then, we call x_j a local center of G on b_j, and furthermore, we have

$$\min_{1 \leq j \leq n} [\, r(x_j)] = r(x_o) \tag{10}$$

Therefore, if the location of the local centers $x_1, x_2, \ldots, x_m$ on branches $b_1, b_2, \ldots, b_m$ are somehow determined, then the point x_o is among the set of points $x_1, x_2, \ldots, x_n$ and is the point in the set for which $r(x_j)$ is a minimum. We will now describe a method for locating the local center x_j on branch b_j of G.

Let us consider the graph of Fig. 2, and the branch $b_j\,(v_p, v_q)$ and an arbitrary vertex v_i in that graph. Let x be an arbitrary point on $b_j\,(v_p, v_q)$, then

$$d(v_i, x) = \min\, [\, d(x, v_p) + d(v_p, v_i),\ d(x, v_q) + d(v_q, v_i)\,] \tag{11}$$

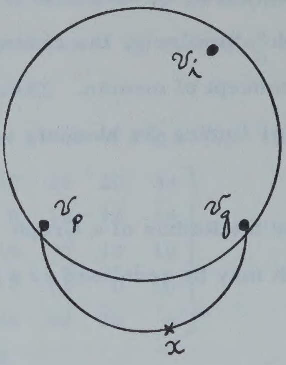


Fig. 2-An illustration of a distance $d(x, v_i)$

Let $d(v_p, v_q) = B_j$ and $d(v_p, x) = x$, then $d(v_q, x) = B_j - x$. In these terms, Eq. (11) may be written as

$$d(v_i, x) = \min [\, x + d(v_p, v_i),\ B_j - x + d(v_q, v_i) \,] \tag{12}$$

Let $f_1(x) = x + d(v_p, v_i)$ and $f_2(x) = B_j - x + d(v_q, v_i)$. The functions $f_1(x)$ and $f_2(x)$ represent straight lines and, therefore, can be plotted easily. These plots are shown in Fig. 3.

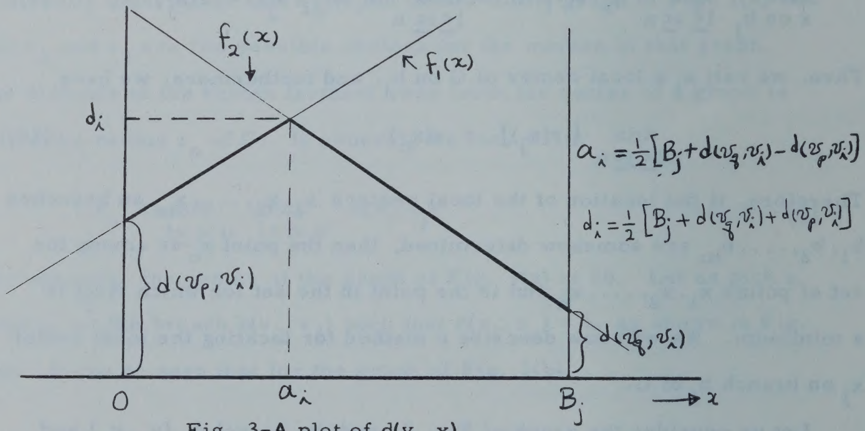


Fig. 3-A plot of $d(v_i, x)$

We would like to obtain a plot of $d(v_i, x) = \min [\, f_1(x),\ f_2(x) \,]$, in the interval $0 \leq x \leq B_j$. Such a plot can be easily obtained from the plots of $f_1(x)$ and $f_2(x)$ and is shown in Fig. 3 by the thick dark line. In Fig. 4, the function $h_i\, d(v_i, x)$ is plotted. Similarly the plots of other functions

$h_1\ d(v_1, x)$, $h_2\ d(v_2, x), \ldots, h_n\ d(v_n, x)$ can be obtained in the interval $0 \leq x \leq B_j$. These functions are all plotted on the same axes as shown in Fig. 4. To find the curve of the function

$$F_j(x) = \max_{1 \leq i \leq n} h_i\ d(v_i, x), \text{ for } x \text{ on } b_j \tag{13}$$

we must pick for each value of x, $(0 \leq x \leq B_j)$, the maximum of all of the function plotted. The curve of the function $F_j(x)$ is also shown in Fig. 4 by the thick dark line. Now, to find the point x_j on $b_j\ (v_p, v_q)$ we must find the point at which $F_j(x)$ is a minimum. Since $F_j(x)$ is a "piecewise linear function," $F_j(s)$ has a finite number of local minima each of which can be easily computed; therefore, there is no difficulty in calculating the position of the point x_j.

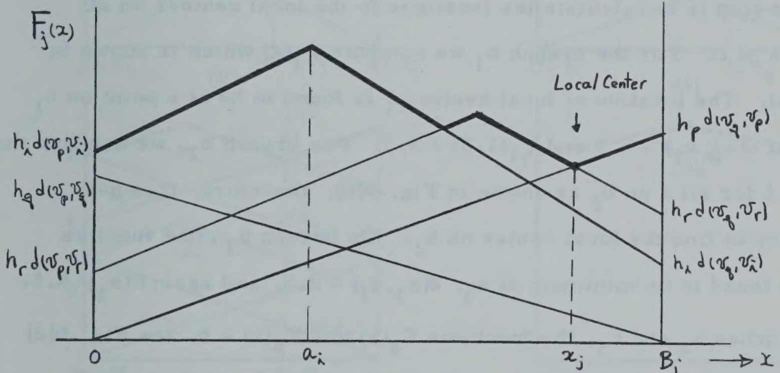


Fig. 4-A representative plot of F(x)

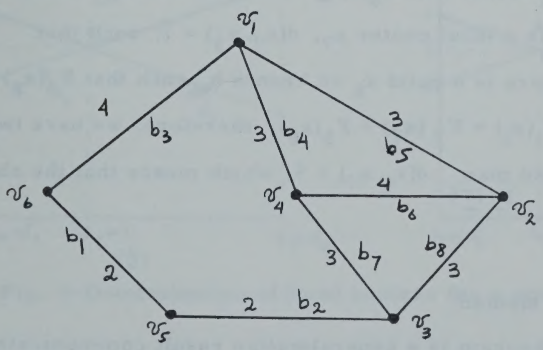


Fig. 5-An example to demonstrate the process for finding the absolute center.

The details of the above process may best be explained by a numerical example. In this example we will assume that weights attached to all vertices is equal to one. Clearly, this assumption does not have much effect on the details of the procedure. Example: Consider the graph G of Fig. 5. We would like to find the location of the absolute center of G.

We first compute the distnace matrix for graph G of Fig. 5.

$$D = \begin{bmatrix} 0 & 3 & 6 & 3 & 6 & 4 \\ 3 & 0 & 3 & 4 & 5 & 7 \\ 6 & 3 & 0 & 3 & 2 & 4 \\ 3 & 4 & 3 & 0 & 5 & 7 \\ 6 & 5 & 2 & 5 & 0 & 2 \\ 4 & 7 & 4 & 7 & 2 & 0 \end{bmatrix} \tag{14}$$

The next step is to calculate the locations fo the local centers on all branches of G. For the branch b_1 we compute $F_1(x)$ which is shown in Fig. 6(a). The location of local center x_1 is found to be at a point on b_1 such that $d(v_6, x_1) = 1.5$ and $F_1(1.5) = 5.5$. For branch b_2, we can find that $F_2(x) \geq 6$ for all x on b_2 as shown in Fig. 6(b); therefore, it is not necessary to find the local center on b_2. On branch b_3, the function $F_3(x)$ is found to be minimum at x_3, $d(x_3, v_1) = 2.5$, and again $F(x_3) = 5.5$. For branches b_4 and b_5, the functions $F_4(x)$ and $F_5(x) = 6$, see Fig. 6(d) and (e). Consider the branch b_6. Since $d(v_6, v_4) = 7$ and $d(v_6 v_2) = 7$, we can immediately decide that $F_6(x) \geq 7$, for all x on b_6. However, on branch b_7 there is a local center x_7, $d(x_7, v_3) = 1$; such that $F_7(x_7) = 5$. Also there is a point x_8 on branch b_8 such that $F_8(x_8) = 5$. Now, clearly $\min F_i(x_i) = F_7(x_7) = F_8(x_8)$; therefore, we have two absolute centers x_7 and x_8 and $\max\limits_{1 \leq i \leq 6} d(v_i, x_7) = 5$, which means that the absolute radius of G is 5.

IV. Absolute Median

The following theorem is a generalization result communicated to the author by A. J. Goldstein [5], who has shown that an absolute

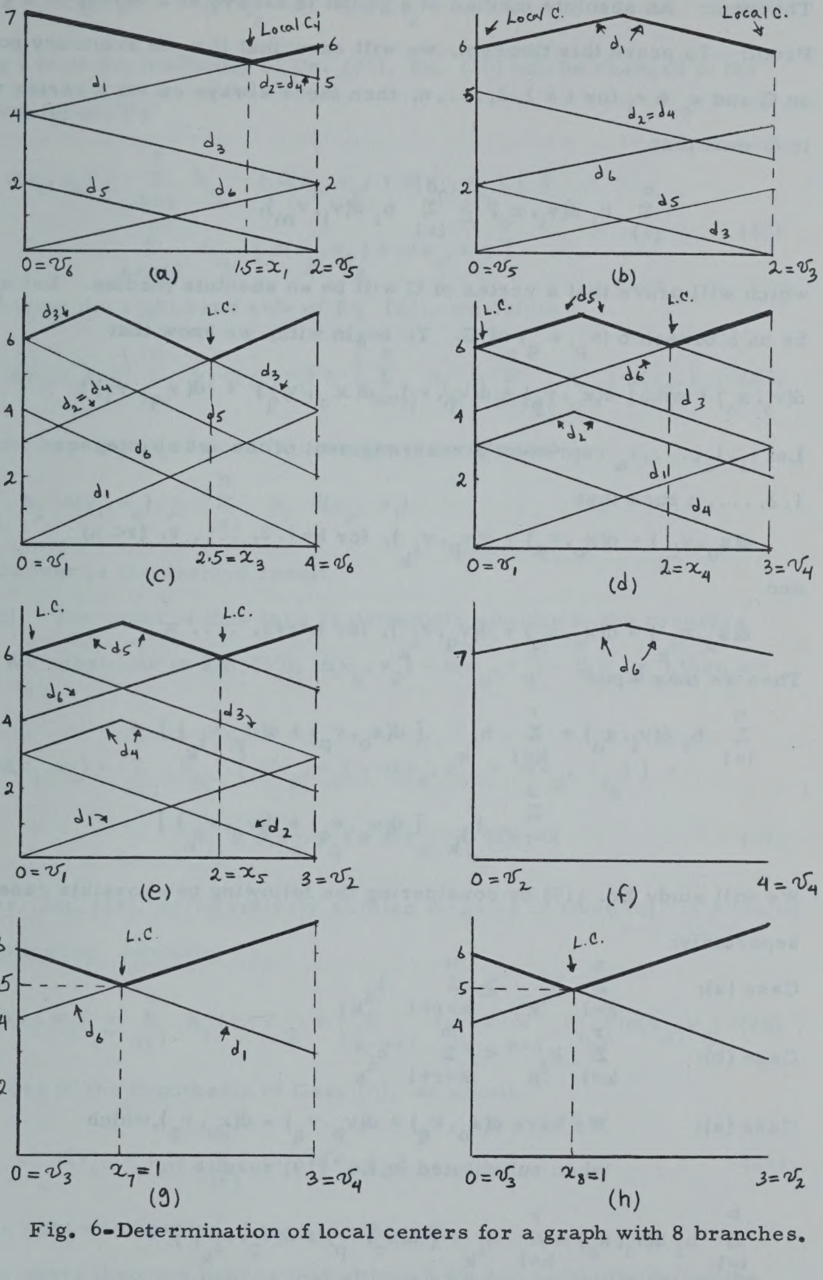


Fig. 6-Determination of local centers for a graph with 8 branches.

median of a connected circuit-less graph (a tree) is always at a vertex.

Theorem: An absolute median of a graph is always at a vertex of a graph.

Proof: To prove this theorem, we will show that if x_o is arbitrary point on G and $x_o \neq v_i$ for $i = 1, 2, \ldots, n$, then there always exists a vertex v_m in G such that

$$\sum_{i=1}^{n} h_i\, d(v_i, x_o) \geq \sum_{i=1}^{n} h_i\, d(v_i, v_m), \tag{15}$$

which will prove that a vertex of G will be an absolute median. Let x_o be on a branch $b(v_p, v_q)$ of G. To begin with, we know that

$$d(v_i, x_o) = \min [\, d(x_o, v_p) + d(v_p, v_i),\ d(x_o, v_q) + d(v_q, v_i)\,] \tag{16}$$

Let $i_1, i_2, \ldots, i_n$ represent a rearrangment of the set of integers $1, 2, \ldots, n$ such that

$$d(x_o, v_{i_k}) = d(x_o, v_p) + d(v_p, v_{i_k}), \text{ for } k = 1, 2, \ldots, r, \ (r \leq n) \tag{17}$$

and

$$d(x_o, v_{i_k}) = d(x_o, v_q) + d(v_q, v_{i_k}), \text{ for } k = r+1, \ldots, n \tag{18}$$

Then we may write

$$\sum_{i=1}^{n} h_i\, d(v_i, x_o) = \sum_{k=1}^{r} h_{i_k} [\, d(x_o, v_p) + d(v_p, v_{i_k})\,] + \sum_{k=r+1}^{n} h_{i_k} [\, d(x_o, v_q) + d(v_q, v_{i_k})\,] \tag{19}$$

We will study Eq. (19) by considering the following two possible cases separately:

Case (a): $\sum_{k=1}^{r} h_{i_k} \geq \sum_{k=r+1}^{n} h_{i_k}$

Case (b): $\sum_{k=1}^{r} h_{i_k} < \sum_{k=r+1}^{n} h_{i_k}$

Case (a): We have $d(x_o, v_q) = d(v_p, v_q) - d(x_o, v_p)$, which when substituted in Eq. (19) results in

$$\sum_{i=1}^{n} h_i\, d(v_i, x_o) = \sum_{k=1}^{r} h_{i_k} [\, d(x_o, v_p) + d(v_p, v_{i_k})\,] + \sum_{k=r+1}^{n} h_{i_k} [\, d(v_p, v_q) - d(x_o, v_p) + d(v_p, v_{i_k})\,] \tag{20}$$

We also know that

$$d(v_q, v_p) + d(v_p, v_{i_k}) \geq d(v_p, v_{i_k}) \tag{21}$$

Making use of the inequality of Eq. (21), Eq. (20) can be changed to the following inequality

$$\sum_{i=1}^{n} h_i\, d(v_i, x_o) \geq \sum_{k=1}^{r} h_{i_k} \left[d(x_o, v_p) + d(v_p, v_{i_k}) \right] + \sum_{k=r+1}^{n} h_{i_k} \left[d(v_p, v_{i_k}) + d(x_o, v_p) \right] \tag{22}$$

Rearranging the right hand side of Eq (22), we obtain

$$\sum_{i=1}^{n} h_i\, d(v_i, x_o) \geq \sum_{i=1}^{n} h_i\, d(v_p, v_i) + \left[\sum_{k=1}^{r} h_{i_k} - \sum_{k=r+1}^{n} h_{i_k} \right] d(x_o, v_p) \tag{23}$$

From the hypothesis of Case (a), we can conclude that

$$\sum_{i=1}^{n} h_i\, d(v_i, x_o) \geq \sum_{i=1}^{n} h_i\, d(v_p, v_i)$$

This however is the desired result.

Case (b)· The proof of this case is extremely similar to the previous case. We substitute in Eq. (19), $d(x_o, v_p) = d(v_p, v_q) - d(v_q, x_o)$, then we obtain

$$\sum_{i=1}^{n} h_i\, d(v_i, x_o) = \sum_{k=1}^{r} h_{i_k} \left[d(v_p, v_q) - d(v_q, x_o) + d(v_p, v_{i_k}) \right] + \sum_{k=r+1}^{n} h_{i_k} \left[d(x_o, v_q) + d(v_q, v_{i_k}) \right] \tag{24}$$

However, Eq. (24), using entirely similar steps as in Case (a), is reduced to the following inequality

$$\sum_{i=1}^{n} h_i\, d(v_i, x_o) \geq \sum_{i=1}^{n} h_i\, (v_i, v_q) + \left[\sum_{k=r+1}^{n} h_{i_k} - \sum_{k=1}^{r} h_{i_k} \right] d(x_o, v_q) \tag{25}$$

Making use of the hypothesis of Case (b), we obtain

$$\sum_{i=1}^{n} h_i\, d(v_i, x_o) > \sum_{i=1}^{n} h_i\, d(v_i, v_q) \tag{26}$$

which is what was desired, hence the theorem.

The above theorem proves that although we can certainly define the term absolute median, but such a point can never be any more of an absolute "median" than a "vertex median."

VI. Conclusions and Further Problems

We have shown that to find the optimum location of a switching center in a communication network, one can limit himself with finding the vertex median of the corresponding graph. However, if one is concerned with building a hospital or a police station on a highway system in such a way that the distance of the farthest community from the police station is a minimum, then one must find the absolute center of the corresponding graph according to the procedure discribed in Section III. We may also look at the problem of the optimum location of the police station in the following way: Suppose h_i represents the average number of daily automobile accidents in community v_i, and furthermore let us assume that a police from the station must visit each scene of accident to make a report. With this sort of situation, the best place to build the station would be at a median of the corresponding graph. The problem becomes more complicated if one is to find the optimum location of the police station with the consideration of a combination of both factors.

The problem of finding the location of the absolute center of a graph may be formulated in terms of game theory. To explain this, we will consider a problem which is slightly more general than the one considered here. Let d(x, y) be the distance between points x and y on G. The person X picks a point x on graph G. If person Y picks a point y on G, then X pays off d(x, y) units of money to Y. X would like to pick a point from which the maximum distance is a minimum. Therefore the least possible amount of money that X can loose is

$$\min_{x \text{ on } G} \max_{y \text{ on } G} d(x, y) = r(x_o) \tag{27}$$

This game theoretic approach seems quite interesting and the author is at the present studying the feasability of such an approach.

REFERENCES

1. O. Ore, "Theory of Graphs", Am. Math. Soc. Colloquium Publication, Providence, R.I., pp. 27-30, 1962.
2. C. Berge, "The Theory of Graphs", Methuen and Co., London, pp. 119-122, 1962.
3. E.F. Moore, "Shortest Path Through a Maze", Proc. Intl. Symposium on Switching Circuits, Harvard Univ., pp. 285-292, April, 1957.
4. S.L. Hakimi and S.S. Yau, "Distance Matrix and the Synthesis of N-port Resistive Networks", Tech. Rep. No. 6, Network Theory Group, Dept. of Elect. Eng., Northwestern Univ., Evanston, Ill.
5. A.J. Goldstein, Private Communication, Bell Telephone Laboratories, Murray Hill, N.J., Summer, 1962.

REGULAR EXPRESSIONS FOR LINEAR SEQUENTIAL CIRCUITS*

J. A. BRZOZOWSKI
Department of Electrical Engineering
The University of Ottawa
Ottawa, Canada

ABSTRACT

This paper considers the class of linear sequential circuits from the regular expression point of view. The circuits studied do not have special starting units which are necessary in the conventional construction of circuits from regular expressions. Since conventional regular expressions are only indirectly related to the circuit structure, a new regular language is developed. Using this language, the regular expression accepted by a linear circuit can be obtained directly from the circuit in a straightforward manner. The regular expressions are then interpreted to provide a word description of the circuit behaviour.

1. Introduction

The behaviour of a finite automaton with a single output z can be characterized by specifying the set of input sequences s resulting in $z = 1$ at the end of s. This set, in general infinite, may be described by a finite formula called a regular expression [1,2]. Given a state diagram of a finite automaton, one can determine[3-6] the corresponding regular expression in a straightforward manner, and there are methods [3,6,7] for obtaining state diagrams from regular expressions.

It is also possible to construct sequential circuits (using AND gates, OR gates, inverters, and unit delays, say) directly from regular expressions[2,4,8]. However, these circuits must be of a special kind[2,4,8], with special starting pulses and the associated circuitry, in addition to the external inputs. If one is given a sequential circuit without starting units, but with a specified initial state, the regular expression cannot be determined directly from the circuit, without finding the state diagram or its equivalent.

In this paper, we restrict our attention to linear sequential circuits. A new regular language, closely related to the circuit structure, will be developed.

It is assumed that the reader is familiar with the basic notions of regular expressions and linear sequential circuits. Suitable background for regular expressions may be found in Ref. 2, and for linear sequential circuits in Refs. 9 or 10.

2. Preliminaries

For simplicity, we shall consider only sequential circuits with a single binary input x and a single binary output z. The values of binary

---

* The research reported here was supported in part by the National Research Council of Canada under grant No. A-1617. Previously available from Dept. of Elec. Engr., The University of Ottawa, Ottawa, Canada as Tech. Rept. No. 63-8, June, 1963.

variables are 0 and 1. Thus, from the regular expression point of view, we shall be concerned with finite sequences of 0's and 1's.

Given two sets P and Q of sequences, we define the following set operations in the usual way:

$P \vee Q$ - Union (Inclusive OR),

$P \& Q$ - Intersection (AND),

$P + Q$ - "Sum" (Exclusive OR),

$\overline{P}$ - Complement (Negation).

These are the usual set or Boolean operations; for example $R = P+Q$ is the set of sequences which are either in P but not in Q, or in Q but not in P, i.e. $P+Q = \overline{P}\&Q \vee P\&\overline{Q}$. All the laws of Boolean algebra hold for these operations.

We also define another operation:

PQ - Concatenation, $PQ = \{s=pq \mid p \in P,\ q \in Q\}$.

The concatenation PQ (P followed by Q) is the set of all sequences obtained by following a sequence $p \in P$ by a sequence $q \in Q$. The concatenation operation is associative, but not commutative, in general. For convenience, PP is denoted by P^2, etc. We also have the distributive laws:

$$(P \vee Q)\,R = PR \vee QR, \tag{2.1}$$

$$R\,(P \vee Q) = RP \vee RQ. \tag{2.2}$$

The empty set of sequences is denoted by ϕ. The set consisting of the sequence of zero length is called λ. We shall not differentiate in the notation for a sequence and for the set consisting of only that sequence; thus λ also denotes the sequence of zero length. The set of all sequences (including λ) is denoted by I. Some of the properties of ϕ, λ and I are:

$$R \vee \phi = R + \phi = R, \tag{2.3}$$

$$R\phi = \phi R = \phi, \tag{2.4}$$

$$R\lambda = \lambda R = R, \tag{2.5}$$

$$\overline{R} = I + R. \tag{2.6}$$

Definition 2.1. A (conventional) regular expression is defined recursively as follows:

1. 0, 1, λ and ϕ are regular expressions.
2. If P and Q are regular expressions, then so are $(P \vee Q)$, (PQ) and P^*.
3. Nothing else is a regular expression, unless its being so follows from a finite number of applications of Rules 1 and 2.

The "star" operation (P^*) is defined by:

$$P^* = \lambda \vee P \vee P^2 \vee \dots \vee P^n \vee \dots \quad . \tag{2.7}$$

Consider Fig. 2.1 showing a linear circuit and its state diagram.

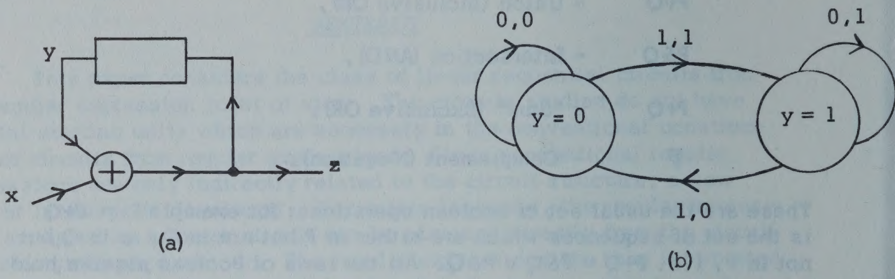


Fig. 2.1. A simple linear circuit. (a) Circuit diagram. (b) State diagram.

The rectangular box is a unit delay, and the circled + is an exclusive OR gate. If the circuit is started in state $y = 0$, then the regular expression from the state diagram[5] can be written in the form:

$$R = (0 \vee 1\, 0^*\, 1)^*\, 1\, 0^* .$$

To the best of our knowledge, this appears to be the simplest (conventional) regular expression which we can write for the given circuit. Notice that the structure of the regular expression appears to be in no way related to the structure of the circuit, although it is related to the state diagram[5]. There is no direct procedure for obtaining such an expression from the circuit.

Observations of this kind have prompted us to re-examine the basic definition of regular expression and to develop a new, more suitable, language to be described. However, before we can proceed, we require some further properties of the set operations.

Definition 2.2 The suffix (prefix) of a sequence s is any final (initial) portion of s, where s and λ are also considered as final (initial) portions of s. (For example, the suffixes of 10110 are: 10110, 0110, 110, 10, 0, and λ.)

Definition 2.3 A set R of sequences is said to have the suffix property (prefix property), if no sequence $s \in R$ is the suffix (prefix) of any other sequence $t \in R$.

Theorem 2.1 If P, Q and R are sets of sequences and

1) If R has the suffix property, then:

a) $(P \& Q)R = PR \& QR,$ (2.8)

b) $(\overline{P} \& Q)R = \overline{PR} \& QR,$ (2.9)

c) $(P + Q)R = PR + QR.$ (2.10)

2) If R has the prefix property, then:

a) $R(P \& Q) = RP \& RQ,$ (2.11)

b) $R(\overline{P} \& Q) = \overline{RP} \& RQ,$ (2.12)

c) $R(P + Q) = RP + RQ.$ (2.13)

Proof: We shall prove the theorem for Case 1, where R has the suffix property; Case 2 follows by an analogous argument.

a) If $s \in (P \& Q)R$, then s can be written in the form $s = ab$, $a \in P$, $a \in Q$ and $b \in R$. Hence $s \in PR$, $s \in QR$ and $s \in PR \& QR$. Thus $(P \& O) R \subset PR \& QR$, where $X \subset Y$ denotes that X is a subset of Y. Now suppose $s \in PR \& QR$; then $s = a_1b_1$, $a_1 \in P$, $b_1 \in R$, and $s = a_2b_2$, $a_2 \in Q$, $b_2 \in R$. But this implies that R contains sequences b_1 and b_2, one of which is a suffix of the other. Here $b_1 = b_2 = b$ is the only allowed possibility, since R has the suffix property. Hence $a_1 = a_2 = a$, and $s = ab$, $a \in P$, $a \in Q$ and $b \in R$. Thus $(P \& Q) R \supset PR \& QR$, and (2.8) follows.

b) If $s \in (\overline{P} \& Q)R$, then $s = a_1b_1$, $a_1 \in \overline{P}$, $a_1 \in Q$ and $b_1 \in R$. Thus $s \in QR$. If $s \in PR$, then $s = a_2b_2$, $a_2 \in P$, $b_2 \in R$. This is only possible if $b_1 = b_2 = b$. But then $a_1 = a_2 = a$, so that $a \in P$ and $a \in \overline{P}$, which is a contradiction. Therefore $s \notin PR$ or $s \in \overline{PR}$, and $(\overline{P} \& Q) R \subset \overline{PR} \& QR$. Conversely, $s \in \overline{PR} \& QR$ implies $s = ab$, $a \in Q$, $b \in R$ and $a \notin P$, or $s \in (\overline{P} \& Q) R$. Hence $(\overline{P} \& Q)R \supset \overline{PR} \& QR$, and (2.9) follows.

c)
$$
\begin{aligned}
(P + Q) R &= (P\&\overline{Q} \vee \overline{P}\&Q) R \\
&= (P\&\overline{Q})R \vee (\overline{P}\&Q)R \\
&= (PR \& \overline{QR}) \vee (\overline{PR} \& QR) \quad \text{(by 1b)}. \\
&= PR + QR.
\end{aligned}
$$

The above laws are distributive laws applicable in limited situations; the distributive law of concatenation over exclusive OR addition will be needed often in later sections.

3. Linear Circuits without Feedback

A sequential circuit is linear[9,10], if it can be realized using only unit delays and exclusive OR gates. We shall assume that the circuits under consideration are not autonomous [9,10]; i.e., the behaviour of the circuits depends non-trivially on the input. It is more convenient to begin by considering linear circuits without feedback. It is assumed throughout that all circuits are started with 0 at the outputs of all delay elements.

An example of a linear circuit without feedback is shown in Fig. 3.1. The circuit operates on the input x(t) to produce an output z(t).

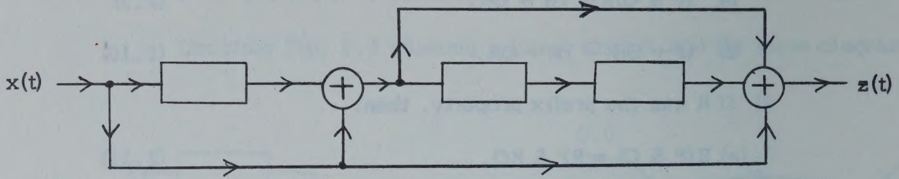


Fig. 3.1. A linear circuit without feedback.

In this case we have:

$$z(t) = (x(t-3) + x(t-2)) + (x(t) + x(t-1)) + x(t)$$

$$= x(t-1) + x(t-2) + x(t-3). \qquad (3.1)$$

One may also view the circuit of Fig. 3.1 as follows: The circuit receives an input sequence and produces a corresponding output sequence. If D denotes the delay operator [9,10] ($D\,x(t) = x(t-1)$), then the input and output sequences are polynomials in D, and are related by the transfer function[9,10]. Let the input sequence be X(D), the output sequence Z(D), and the transfer function H(D); then:

$$Z(D) = H(D)\,X(D). \qquad (3.2)$$

For the given circuit, the transfer function is, by inspection:

$$H(D) = (D + 1)(D^2 + 1) + 1$$

$$= 1 + 1 + D + D^2 + D^3 = D + D^2 + D^3. \qquad (3.3)$$

Alternately, if the input sequence is considered as a polynomial in D, then the circuit multiplies[9] that polynomial by a fixed polynomial H(D).

It is clear that the output of a linear circuit without feedback, and having n delay elements, is a linear combination of the present value and of the past n values of the input. Thus we can represent any linear circuit without feedback by the canonical form shown in Fig. 3.2, where the a_j are either open circuits (0) or short circuits (1). The transfer function for this type of circuit is a polynomial in D with coefficients in the field[11] $GF(2) = \{0,1\}$.

Definition 3.1 A linear circuit without feedback will be called a linear definite circuit. (Such a circuit realizes a definite event [1,7].)

If we define the sum of two definite linear circuits M_1 and M_2 as in Fig. 3.3(a), and their concatenation or product as in Fig. 3.3(b), it is clear that the resulting circuits are also definite linear. It is also clear

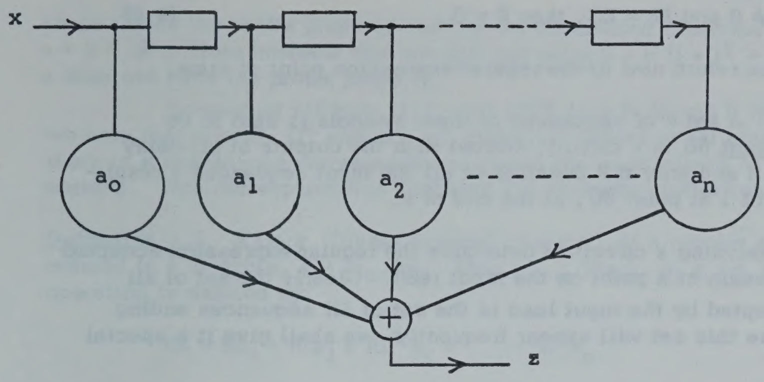
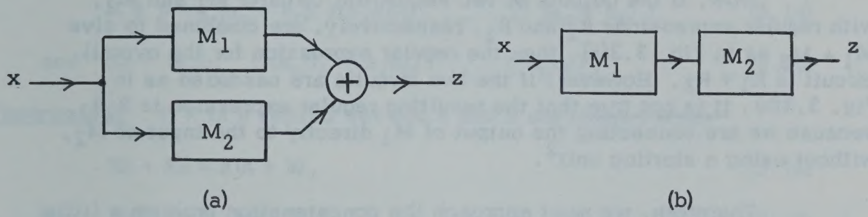


Fig. 3.2. Canonical form of linear circuit without feedback.



(a) (b)

Fig. 3.3 Operations on circuits. (a) Sum. (b) Product.

that, if H_1 and H_2 are the transfer functions for M_1 and M_2, respectively, then the transfer functions of $M_1 + M_2$ and M_1M_2 are $H_1 + H_2$ and H_1H_2, respectively. The algebra of linear definite circuits is isomorphic to the algebra of the corresponding transfer functions because of the one-to-one correspondence one can establish between the circuit structure and the form of the transfer function.

Theorem 3.1 The set of linear definite circuits or the set of their transfer functions is an integral domain[11] isomorphic to the integral domain of polynomials with coefficients in GF(2), under addition and multiplication as defined above.

Proof: The properties[11] of an integral domain are easily verified. If E, F and G are transfer functions of linear definite circuits, then:

$$E + F = F + E;\ EF = FE\ , \tag{3.4}$$

$$E + (F + G) = (E + F) + G;\ E\,(FG) = (EF)\ G\,, \tag{3.5}$$

$$E\,(F + G) = EF + EG\,, \tag{3.6}$$

$$E + 0 = E;\ E1 = E\,, \tag{3.7}$$

$$E + E = 0 \text{ (additive inverse)}\,, \tag{3.8}$$

If $E \neq 0$ and $EF = EG$, then $F = G$. (3.9)

Let us return now to the regular expression point of view.

Definition 3.2 A set R of sequences of input symbols is said to be accepted at a point (R) in a circuit, started with the outputs of all delay elements at 0, if and only if R consists of all the input sequences s resulting in a signal of 1 at point (R), at the end of s.

In analyzing a circuit to determine the regular expression accepted at a point, we begin at a point on the input lead. Clearly the set of all sequences accepted by the input lead is the set of all sequences ending in a 1. Because this set will appear frequently, we shall give it a special symbol u:

$$u = I1. \tag{3.10}$$

Now, if the outputs of two sequential circuits M_1 and M_2, with regular expressions R_1 and R_2, respectively, are combined to give $M_1 + M_2$ as in Fig. 3.3(a), then the regular expression for the overall circuit is $R_1 + R_2$. However, if the two circuits are cascaded as in Fig. 3.3(b), it is not true that the resulting regular expression is R_1R_2, because we are connecting the output of M_1 directly to the input of M_2, without using a starting unit[2].

Therefore, we must approach the concatenation problem a little differently; let us consider the effect of a unit delay at the output of a circuit, as shown in Fig. 3.4. The resultant expression is $R(0 \vee 1)$; i.e.,

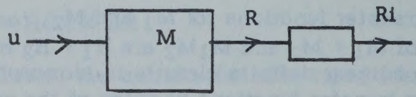


Fig. 3.4. The effect of a unit delay.

the present input is unimportant and can be either 0 or 1. The expression $0 \vee 1$ will be given a special symbol i, because of its frequent use:

$$i = 0 \vee 1 = 0 + 1. \tag{3.11}$$

(Note that if $P \& Q = \emptyset$, then $P + Q = P \vee Q$).

With the above results we can analyze any linear definite circuit for regular expressions by inspection. Consider the circuit of Fig. 3.1. Beginning at the input and proceeding towards the output, we find:

$$\begin{aligned} R &= (u + ui)\, i^2 + (u + ui) + u \\ &= ui + ui^2 + ui^3, \end{aligned} \tag{3.12}$$

since i^2 has the prefix property and can be distributed (Theorem 3.1), and $u + u = \phi$. Note however that we may not write $R = u(i + i^2 + i^3)$ because u does not have the prefix property.

Comparing $z(t)$ from (3.1) with $H(D)$ from (3.3) and R from (3.12), we note that $x(t-n)$, D^n and ui^n are either all absent or all present, i.e., there is a one-to-one correspondence among the coefficients of all three objects. This correspondence suggests the following definition:

Definition 3.3 A linor (linear integral operator) on a regular set R is an ordered (n + 1)-tuple, $A = (a_o, ia_1, i^2a_2, \ldots, i^na_n)$, $a_j \in \{\phi, \lambda\}$, whose operation is defined by:

$$RA = Ra_o + Ria_1 + Ri^2 a_2 + \ldots + Ri^n a_n. \tag{3.13}$$

Definition 3.4 The sum and product of two linors A and B is defined in a way analogous to that of sum and product of polynomials, i.e.:

$$A + B = ((a_o + b_o), i(a_1+b_1), \ldots, i^n (a_n + b_n)), \tag{3.14}$$

and $$AB = (a_ob_o, i(a_ob_1 + a_1b_o), \ldots, i^na_nb_n). \tag{3.15}$$

Theorem 3.2 If R is a regular set and A and B are linors, then:

$$RA + RB = R(A + B), \tag{3.16}$$

and $$(RA)B = R(AB). \tag{3.17}$$

Proof: Since $Ri^ka_k + Ri^kb_k = Ri^k(a_k + b_k)$, for $a_k, b_k \in \{\phi, \lambda\}$, as can be easily verified, (3.16) follows. For the second part,

$(RA)B = (RA)b_o + (RA)ib_1 + \ldots + (RA)i^nb_n$, by definition. For a typical term we have:

$$(RA)i^kb_k = (Ra_o + Ria_1 + \ldots + Ri^na_n)i^kb_k.$$

Now i^kb_k is either ϕ or i^k; in either case it has the suffix property and can be distributed. Furthermore, for $a_j, b_k \in \{\phi, \lambda\}$, their position in the concatenation is unimportant, i.e. $Ri^ja_ji^kb_k = Ri^{j+k}a_jb_k$. Adding such terms over j and k, we will obtain (3.17).

Theorem 3.3 Every regular expression describing a linear definite circuit can be written in the form:

$$R = uA, \tag{3.18}$$

where A is a linor. Furthermore, the set of all linors forms an integral domain isomorphic to the domain of polynomials with coefficients in G F (2).

Proof: From the canonical form of Fig. 3.2 we can write the expression in the form

$$R = ua_o + uia_1 + \ldots + ui^na_n = uA.$$

Also, we establish a one-to-one correspondence between the coefficients of D^k in the transfer function and the coefficient of ui^k above, where 0 corresponds to ϕ and 1 to λ. The correspondence of sums and products is clearly preserved (Definition 3.4). Theorem 3.2 ensures that the sum and product operations are meaningful for regular expressions.

It can be seen that, although regular expressions for linear definite circuits do not themselves form an integral domain, each such expression can be put in the form $R = uA$, where A is a linor, and linors form an integral domain. In other words we may restate the integral domain properties for regular expressions as follows: The circuit performs a linear operation on the set u of sequences at the input lead; to add two circuits, add the corresponding linors; to multiply two circuits, multiply the corresponding linors. The properties given in (3.4) -(3.9) can be restated for regular expressions: If A, B, C are linors, then;

$$u(A + B) = u(B + A);\ u(AB) = u(BA),$$

$$u((A + B) + C) = u(A + (B + C));\ uA(BC) = u(AB)C,$$

$$uA(B + C) = u(AB + AC),\ \text{etc.}$$

4. The Effects of Feedback

As yet we have not discussed the possibility of feedback. The addition of a feedback construction to the delay and sum operations will permit us to study the entire class of linear circuits.

In the simple feedback circuit of Fig. 2.1, we cannot write the regular expression R, accepted by the circuit, directly from the circuit. However we can obtain, by inspection, the equation which R must satisfy:

$$R = Ri + u.$$

If we group the terms involving R on the left and those involving u on the right, we have:

$$R + Ri = u. \qquad (4.1)$$

Using our linor notation, let $A = (\lambda)$ and $B = (\lambda, i)$. Then the form of (4.1) is:

$$RB = uA. \qquad (4.2)$$

The transfer function for the circuit is $\frac{1}{1 + D}$, which suggests that the solution of (4.2) should be $R = u(\frac{A}{B})$, provided that division of linors can be well defined. Since linors form an integral domain, the introduction of $\frac{A}{B}$ is analogous to the construction of the rationals from the integers[11]. Unfortunately, the analogy with circuits is not complete because transfer functions like $\frac{1}{D}$ are not physically realizable because we cannot obtain an ideal predictor. Let us consider, however, functions which are ratios of polynomials in D. Each such function can be expressed in the form:

$$F(D) = \frac{D^p\,(1 + a_1D + \ldots + a_mD^m)}{D^r\,(1 + b_1D + \ldots + b_nD^n)}$$

$$= \frac{(1 + a_1D + \ldots + a_mD^m)\;D^{p-r}}{(1 + b_1D + \ldots + b_nD^n)}. \qquad (4.3)$$

If $(p-r) \geqslant 0$, then F(D) is realizable and the corresponding circuit can take the form of the circuit in Fig. 4.1. On the other hand, if $(p-r) < 0$, then F(D)

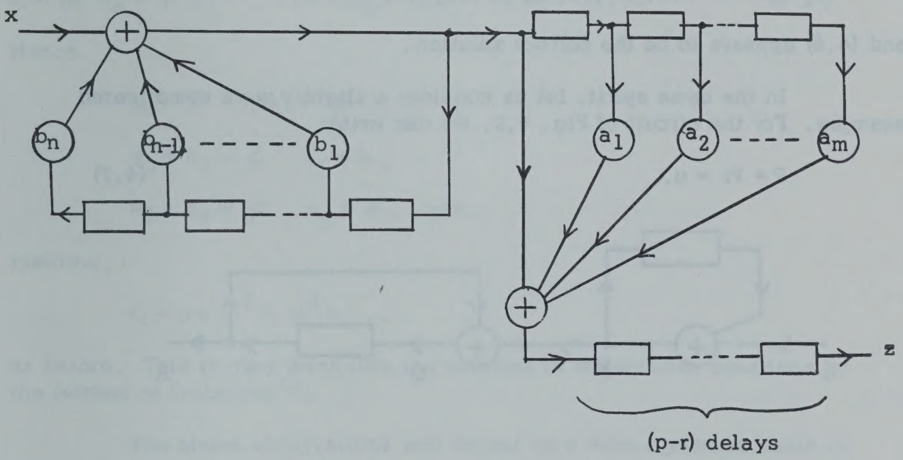


Fig. 4.1. Realization of F(D).

is not realizable. However, we may interpret F(D) in this case as realizable with a delay of (r-p) units; i.e., $F(D)\,D^{r-p}$ is realizable. It is easily verified that the set of transfer functions realizable with some delay is a field[11], since multiplicative inverses are now defined. We could, by analogy solve regular equations of the form (4.2) by defining inverses of linors; however, regular expressions obtained in this way appear to be rather difficult to interpret, and we shall use a different approach.

Returning to the transfer function of the circuit in Fig. 2.1, it should be pointed out that D has no numerical value and that $\frac{1}{1+D}$ is used simply as a short form for an infinite sum;

$$\frac{1}{1+D} = 1 + D + D^2 + \dots + D^n + \dots \quad . \tag{4.4}$$

By analogy, suppose we expand the operator $\frac{A}{B}$ in a similar way:

$$\frac{(\lambda)}{(\lambda, i)} = (\lambda, i, i^2, \dots, i^n, \dots) \quad . \tag{4.5}$$

We obtain a linor with an infinite number of components, suggesting that the solution of (4.1) should be:

$$R = u + ui + ui^2 + \dots + ui^n + \dots \quad . \tag{4.6}$$

Forgetting about rigour for the moment, let us substitute (4.6) into (4.1). Then we have

$$(u + ui + \dots) + (ui + ui^2 + \dots) = u,$$

and (4.6) appears to be the correct solution.

In the same spirit, let us consider a slightly more complicated example. For the circuit of Fig. 4.2, we can write:

$$P + Pi = u, \tag{4.7}$$

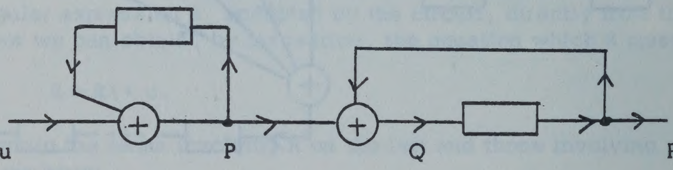


Fig. 4.2. Circuit C.

$$Q + Qi = P \tag{4.8}$$

$$R = Qi, \tag{4.9}$$

where P and Q are the regular expressions at the inputs to the delay elements. We may eliminate P as follows:

$$Pi = P + u = Qi^2 + Qi,$$

$$Qi^2 + Qi = P + u = Qi + Q + u,$$

or $$Q + Qi^2 = u. \tag{4.10}$$

The solution suggested by the analogy with the transfer function $\frac{1}{1+D^2}$ is

$$Q = u + ui^2 + ui^4 + \dots, \tag{4.11}$$

and $$R = ui + ui^3 + ui^5 + \dots. \tag{4.12}$$

Alternately, we can obtain the same answer by assuming that Q in (4.10) has the solution:

$$Q = ua_o + uia_1 + ui^2a_2 + \dots, \tag{4.13}$$

and substituting in (4.10) to evaluate the unknown coefficients a_j. Thus

$$u = (ui^2a_o + ui^3a_1 + \dots) + (ua_o + uia_1 + ui^2a_2 + \dots). \tag{4.14}$$

Hence

$$a_o = \lambda \quad , a_1 = \phi \ ,$$

$$a_o + a_2 = \phi \ , a_2 = \lambda,$$

$$a_1 + a_3 = \phi \ , a_3 = \phi \ , \text{etc.},$$

yielding,

$$Q = u + ui^2 + ui^4 + \dots,$$

as before. This is very much like the solution of differential equations by the method of Frobenius[12].

The above observations will be put on a more rigourous basis in the next section.

5. Linear Regular Expressions

We are concerned with analyzing a linear circuit to determine the set of sequences accepted by that circuit. Let H(D) be the transfer function of the form (4.3), with denominator polynomial $B(D) = 1 + b_1 D + \dots + b_n D^n$. Let $1 + D^s$ be the polynomial of lowest degree such that B(D) is a factor of $1 + D^s$, i.e.,

$$1 + D^s = B(D)\ C(D). \tag{5.1}$$

One can always find an $s \geqslant n$ satisfying this condition[9]. Then $H(D)$ can be written in the form:

$$H(D) = \frac{G(D)\, D^{p-r}}{1 + D^s}, \tag{5.2}$$

by multiplying the numerator and the denominator by $C(D)$. The polynomial $G(D) = 1 + g_1 D + \ldots g_q D^q$, is of degree $q = m + s - n$. The circuit corresponding to $H(D)$ from (5.2) consists of a single feedback loop with s unit delays followed by a circuit without feedback, as shown in Fig. 5.1. The set R of sequences accepted by the circuit is given by:

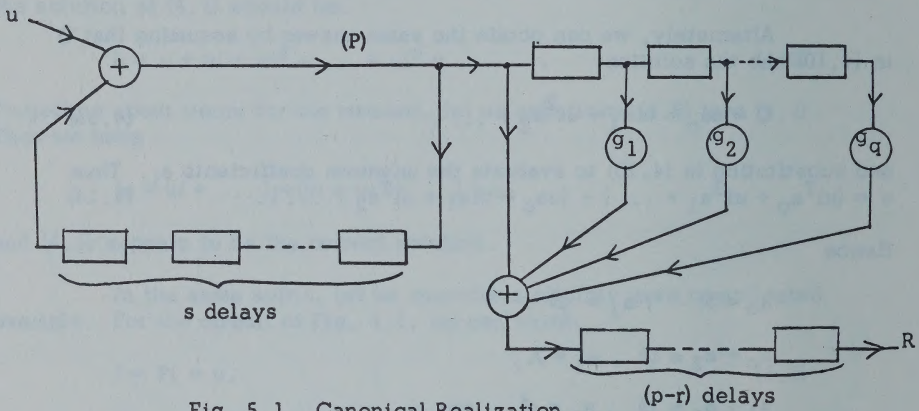


Fig. 5.1. Canonical Realization

$$R = PG\, i^{p-r}, \tag{5.3}$$

where G is the linor corresponding to $G(D)$ in (5.2) and P is the solution of:

$$P + Pi^s = u. \tag{5.4}$$

Thus the main problem is the problem of the feedback connection or the solution of (5.4).

Definition 5.1. A feedback operator of order s is denoted by $\widehat{i^s}$, and is such that $u\widehat{i^s}$ is the solution of (5.4) for $s > 0$, and $\widehat{i^0} = \lambda$.

Theorem 5.1. The set of sequences accepted by any 1-input, 1-output linear circuit can be expressed in the form:

$$R = u\,\widehat{i^s}\; L, \tag{5.5}$$

where L is a linor.

Proof: If the circuit has no feedback then it must perform a linor operation on u, i.e., $R = uL = u \lambda L = u\widehat{i^0} L$, as in Section 3. If there is feedback, let $L = G i^{p-r}$; then R is of the form (5.5).

To illustrate the above ideas consider the circuit of Fig. 5.2. The output set R can be expressed in terms of P, for $R = Pi^2$. Now P must satisfy:

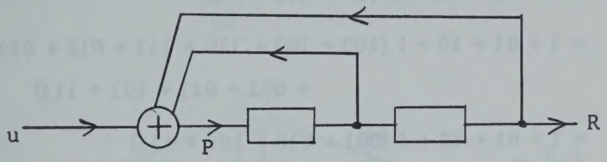


Fig. 5.2. Circuit M.

$$P + Pi + Pi^2 = u. \qquad (5.6)$$

Multiplying (5.6) by i on the right and adding the result to (5.6) we have

$$(P + Pi + Pi^2) + (Pi + Pi^2 + Pi^3) = u + ui,$$

or

$$P + Pi^3 = u + ui = u(\lambda, i). \qquad (5.7)$$

The solution of $X + Xi^3 = u$ is $X = u\widehat{i^3}$, which must then be multiplied by the linor (λ, i) to give: $P = u\widehat{i^3}(\lambda, i) = u\widehat{i^3} + u\widehat{i^3}\, i$. Finally

$$R = Pi^2 = u\widehat{i^3}\, i^2 + u\widehat{i^3}\, i^3, \qquad (5.8)$$

is the required regular expression.

We have shown that a regular expression, though not of the conventional type, can be obtained for every linear circuit by straightforward manipulations of its transfer function (or, equivalently, of the equations which the regular sets must satisfy). The process does not require the construction of a state diagram. We have defined our regular operators in a new way to fit the nature of the linear circuit. Such regular expressions will be more useful, if we can find a meaningful interpretation of the operators. We proceed to discuss this next.

6. Interpretation

If the circuit is definite (contains no feedback), then the translation from our form to a conventional form is straightforward. Suppose we obtain from a circuit the expression $R = u + ui + ui^2$. Realizing that

$$ui^2 = I\,1\,i^2 = I\,(100 + 101 + 110 + 111),$$

$$ui = I\,1\,i = 10 + 11 + I\,(i1i),$$

$$u\; = I\,1 = 1 + 01 + 11 + I\,(i^2 1),$$

we have:

$$R = 1 + 01 + 10 + I\,(1i^2 + i1i + i^2 1)$$

$$= 1 + 01 + 10 + I\,(100 + 101 + 110 + 111 + 010 + 011 + 110 + 111$$
$$+ 001 + 011 + 101 + 111)$$

$$= 1 + 01 + 10 + I\,(001 + 010 + 100 + 111)$$

$$= 1 \vee 01 \vee 10 \vee I\,(001 \vee 010 \vee 100 \vee 111),$$

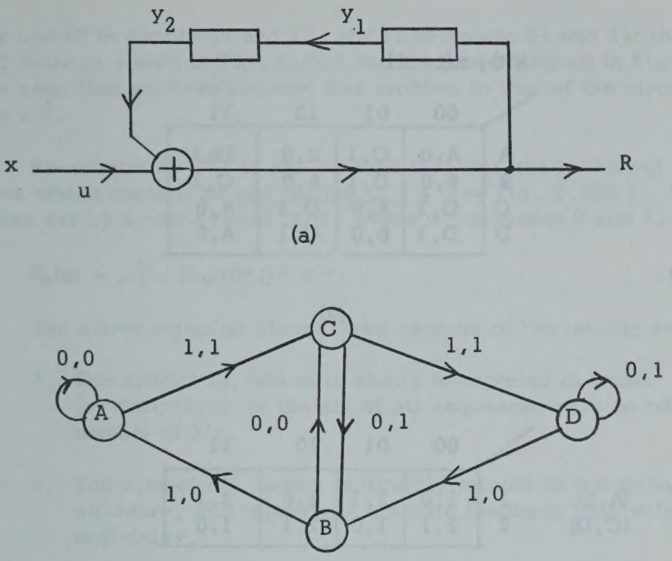
which is a conventional expression in definite [2,7] form, and contains the sequences 1, 01, 10 and all sequences ending in 001, 010, 100, or 111.

When the circuit does contain feedback, the translation is not quite as straightforward. For example, for the circuit of Fig. 2.1, realizing $R = u\hat{i}$ in our terminology, we must find some conventional form such as $R = (0 \vee 10^*\,1)^*\,10^*$. Since it appears unlikely that a direct procedure can be found for such a translation, we adopt a different approach.

If we regard our linear regular expressions as a new language, then we must learn the vocabulary of that language, including certain "idiomatic" phrases, before we can attempt to understand more complicated statements in that language. Fortunately, not too much is required, because the essential symbol to be understood is the "hat" operator $\widehat{i^s}$, in view of Theorem 5.1.

Let us begin with an example. The circuit of Fig. 6.1 (a) realizes $R = u\widehat{i^2}$. Its state diagram and state table are shown in Figs. 6.1 (b) and (c), respectively.

The delay in the feedback loop is 2 units; this suggests the examination of sequences of length 0, 2, 4, ... , i.e., even length sequences. At time $t = 0$, we start in state A ($y_1 = 0$, $y_2 = 0$), and we are interested in the state at $t = 2$. This can be represented best by a state table in which the inputs are not 0 and 1, but sequences 00, 01, 10 and 11.



(a)

(b)

		Input	
$y_1\ y_2$	Present State	0	1
0 0	A	A, 0	C, 1
0 1	B	C, 1	A, 0
1 0	C	B, 0	D, 1
1 1	D	D, 1	B, 0

Next state, output

(c)

Fig. 6.1. Realization of $u\hat{i}^2$. (a) Circuit. (b) State diagram. (c) State table.

The table is shown in Fig. 6.2 (a). We note that the table can be reduced to the table of Fig. 6.2(b), because A is equivalent to B and C to D. Next note that the next state entries and output entries depend only on the last input: Columns 00 and 10 are identical, and so are 01 and 11. Let us x(t), x(t + 1)

	00	01	10	11
A	A,0	C,1	B,0	D,1
B	B,0	D,1	A,0	C,1
C	C,1	A,0	D,1	B,0
D	D,1	B,0	C,1	A,0

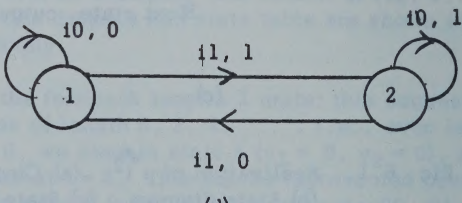
(a)

		00	01	10	11
(A,B)	1	1,0	2,1	1,0	2,1
(C,D)	2	2,1	1,0	2,1	1,0

(b)

	i0	i1
1	1,0	2,1
2	2,1	1,0

(c)



(d)

Fig. 6.2. (a) Table for sequences of length 2. (b) Reduction of rows. (c) Reduction of columns. (d) State diagram.

therefore use i0 to denote 00 and 10, and i1 to denote 01 and 11; the resulting table is shown in Fig. 6.2(c) and the state diagram in Fig. 6.2(d). It can be seen that we have reduced this problem to that of the circuit (Fig. 2.1) realizing $u\widehat{i}$.

Re-examining $u\widehat{i}$, it can be seen that this set is the set of sequences which contain an odd number of 1's (see Fig. 2.1(b)). Let us denote this set by a new symbol $f_o(x)$, where x represents 0 and 1, i.e.,

$$f_o(x) = u\widehat{i} = (0 \vee 10^*1)^*\, 10^*. \tag{6.1}$$

The above equation shows three aspects of the regular set:

1. The symbol f_o, which is easily interpreted in words (by definition) as the set of all sequences with an odd number of 1's.

2. The symbol $u\widehat{i}$, which is directly related to the circuit structure, and represents a single feedback loop with a unit delay.

3. The conventional regular expression, closely related to the state diagram.

In our language the symbol $u\widehat{i}$ must be accepted as basic and its meaning is clear.

For the circuit of Fig. 6.1 (a), let Re denote the sequences of even length which are accepted by the circuit. From the diagram of Fig. 6.2(d),

$$R_e = (i0 \vee i1(i0)^*\, i1)^*\, i1(i0)^* = f_o\,(ix), \tag{6.2}$$

where $f_o(ix)$ is the regular expression (6.1) with each x replaced by ix (0 by i0, 1 by i1). Now R_e can be interpreted as the set of sequences of even length which have an odd number of 1's at times $t = 0,2,4, \ldots$; the inputs at odd times are disregarded.

It is still necessary to find the set of sequences of odd length accepted by the circuit of Fig. 6.1(a). The table of Fig. 6.2(a) still applies, except that the starting state now is not necessarily A, but it is the state reached from A after one unit of time. This state can be either A, if a 0 is first received or C, if a 1 is first received. If it is A, then the circuit accepts the set 0 $f_o(ix)$; if it is C, the circuit accepts $1\,\overline{f_o(ix)} = 1\, f_e(ix)$, where $f_e(x)$ is the set of sequences which have an even number of 1's and it is the complement of $f_o(x)$. Summing up, we can write

$$R = u\widehat{i^2} = f_o(ix) + 0\, f_o(ix) + 1\, f_e(ix).$$

This then is a meaningful interpretation of $\widehat{i^2}$.

Let the set of all sequences of length $\leqslant k$ be denoted by J_k.

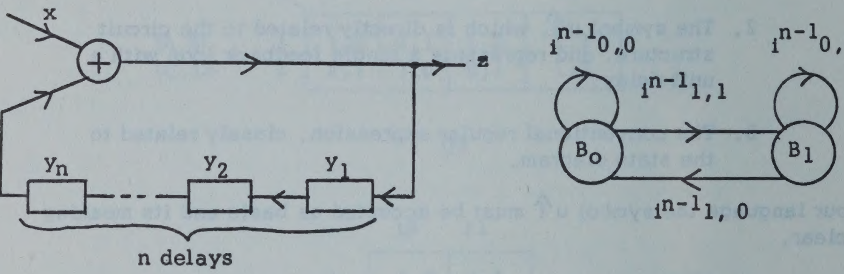
$$J_k = \lambda + i + i^2 + \ldots + i^k. \tag{6.3}$$

Then we can prove the following result:

Theorem 6.1

$$u\widehat{i^n} = J_{n-2}\,(0\, f_o\,(i^{n-1}\,x) + 1\, f_e\,(i^{\,n-1}x)\,) + I_o\,(i^{n-1}x). \tag{6.4}$$

Proof: The circuit under consideration is that of Fig. 6.3 (a). If y_j denotes the output of delay j at time t and y_j'' denotes the next value of y_j, then



(a) (b)

Fig. 6.3. Illustrating $R = u\,\widehat{i^n}$. (a) Circuit (b) State diagram.

$$y_n' = y_{n-1}\,,$$

$$y_{n-1}' = y_{n-2}\,,$$

$$y_2' = y_1\,,$$

$$y_1' = x + y_n. \tag{6.5}$$

The states of the circuit can be partitioned into two blocks: the block Bo with $y_1 = 0$ and the block B_1 with $y_1 = 1$. The sequences of length n can be partitioned into those in $S_o = i^{n-1}0$ and those in $S_1 = i^{n-1}1$. As far as sequences of length kn are concerned, where k is an integer $\geqslant 0$, the behaviour of the circuit can be described by the state table of Fig. 6.3 (b). This follows from (6.5) as in the example for $n = 2$, discussed above. Hence sequences of length kn which are accepted by the circuit are given by:

$$u\hat{i}^{n} = f_0\,(i^{n-1}x)\,. \tag{6.6}$$

If the length of the sequence is $m + kn$, where $1 \leq m \leq n-1$, we consider the first m symbols of the sequence to establish the starting state and then use the state diagram of Fig. 6.3(b). It is easily verified that, starting in A, all sequences in $J_{n-2}0$ return the graph to a state in B_0, and those in $J_{n-2}1$, send the graph to a state in B_1. From then on, the remainder of the sequence has length kn, and the state diagram of Fig. 6.3(b) applies. Hence the theorem follows.

The use of Theorem 6.1 and linors allows us to analyze any 1-input, 1-output linear circuit. For example, if

$$R = u\hat{1} + u\hat{1}\,i = f_0(x) + f_0(x)\,i$$

$$= f_0(x)\ \&\ \overline{f_0(x)i}\ \vee\ \overline{f_0(x)}\ \&\ (f_0(x)\ i)\ ,$$

then R contains a sequence s of length k if s has an odd number of 1's but its prefix of length k-1 does not have an odd number of 1's or if s has an even number of 1's ($s \in \overline{f_0(x)}$) but its prefix of length k-1 has an odd number of 1's. It is seen that, basically, a linear circuit accepts or rejects a sequence on the basis of the parity of 1's in that sequence.

7. Concluding Remarks

We have described a new regular language applicable to linear sequential circuits. To the best of our knowledge, this is the first such language which permits the direct treatment of cascaded circuits, without starting units, and which is very closely related to the circuit structure and transfer function. Moreover, an attempt has been made to give meaning to the expressions so that they may also be useful for synthesis of circuits from word statements. It is hoped that the extension of this approach to the case of general sequential circuits will provide further insight into the problems of sequential circuits.

REFERENCES

1. S.C. Kleene, "Representation of events in nerve nets and finite automata", in "Automata Studies, Annals of Math. Studies", C.E. Shannon and J. McCarthy, Eds., Princeton University Press, Princeton, N.J., No. 34, pp. 3-41; 1956.

2. J.A. Brzozowski, "A survey of regular expressions and their applications", IRE Trans. on Electronic Computers, vol. EC-11 pp. 324-335; June, 1962. (Also Dept. of Elec. Engrg., Digital Systems Lab., Princeton University, Princeton, N.J., Tech. Rept. No. 4; April, 1961.)

3. R. McNaughton and H. Yamada, "Regular expressions and state graphs for automata", IRE Trans. on Electronic Computers, vol. EC-9, pp. 39-47; March, 1960.

4. D.N. Arden, "Delayed logic and finite state machines", Proc. 2nd Ann. Symp. on Switching Circuit Theory and Logical Design, Chicago, Ill., pp. 133-151; September, 1961. (Also in "Theory of Computing Machine Design", University of Michigan Press, Ann Arbor, Mich., pp. 1-35; 1960.)

5. J.A. Brzozowski and E.J. McCluskey, Jr., "Signal flow graph techniques for sequential circuit state diagrams", IEEE Trans. on Electronic Computers, vol. EC - 12; April, 1963. (Also Dept. of Elec. Engrg., Digital Systems Lab., Princeton University, Princeton, N.J., Tech. Rept. No. 5; April, 1961.)

6. J.A. Brzozowski, "Properties of Regular Expressions and State Diagrams", Dept. of Elec. Engrg., Princeton University, Princeton, N.J., Tech. Rept. No. 15; March, 1962.

7. J.A. Brzozowski, "Canonical Regular Expressions and Minimal State Graphs for Definite Events", Dept. of Elec. Engrg., Digital Systems Lab., Princeton University, Princeton, N.J., Tech. Rept. No. 22; April, 1962. (Also to be published in Proc. Symp. on The Mathematical Theory of Automata, Polytechnic Institute of Brooklyn, Brooklyn, N.Y.; April, 1962.)

8. I.M. Copi, C.L. Elgot, and J.B. Wright, "Realization of events by logical nets", J. Assoc. Comp. Mach., vol. 5, pp. 181-196; April, 1958.

9. W.W. Peterson, "Error-Correcting Codes," The M.I.T. Press and John Wiley and Sons, Inc., New York, N.Y.; 1961.

10. W.A. Davis, "An Introduction to Linear Sequential Machines", Dept. of Elec. Engrg., The University of Ottawa, Ottawa, Canada, Tech. Rept. No. 63-2; February, 1963.

11. G. Birkoff and S. MacLane, "A Survey of Modern Algebra", The MacMillan Company, New York, N.Y.; 1941.

12. H.T.H. Piaggio, "Differential Equations", G. Bell and Sons Ltd., London, 1956.

ON TOPOLOGICAL FORMULAE*

SAMUEL C. LEE
Department of Electrical Engineering
University of Illinois
Urbana, Illinois

Abstract:

A complete table containing the trees of thirty-four basic graphs is given which can be used for finding any network function of a finite graph. In this paper the graph is classified according to its nullity and rank. For a network whose nullity is less than six (if it is six, the rank is not greater than six), the network function can be directly obtained by using the table. For the nullity of a network which is not much higher than six, a method of obtaining the network function of an n-node graph with the help of the trees of its (n-1)-node subgraph, the trees already being known, is introduced. Finally a technique of "graph decomposition" is given for finding the network function for high-nullity and high-rank networks.

I. Introduction

The foundation of modern network analysis is the two fundamental Kirchhoff's laws published in 1845. By using the Kirchhoff voltage and current laws, the current and the voltage of any branch of an electric network can be found, and also the network function of a given network can be obtained. However, determining a network function of high-nullity and high-rank network involves solving a set of fairly large number of simultaneous equations which is quite cumbersome.

Recently network theorists[1-6] have investigated the relation existing between the network function and network topology, and have shown that the network function can be expressed topologically in terms of trees and two-trees. By using these topological formulas for high-nullity and high-rank networks, the network function can be written explicitly and easily and

---

*This work was supported by the Air Force Office of Scientific Research under Grant No. AF-AFOSR 177-63.

a great deal of labor saved. It is obvious that it is much simpler to find the trees and two-trees from the network topology than solving the simultaneous equations.

This paper presents several methods for finding any network function in topological terms for a finite electric network. First, a complete table is provided which contains the trees of thirty-four basic graphs. The table can be used to find any network function of a network whose nullity is less than six. If the nullity is six, the table can be used to find any network function of a network whose rank is six or less. For the higher nullity and rank networks, a general succinct method is given for finding the topological formulas of an n-node network from the trees of its (n-1)-node subgraph, the trees already being known. Finally, a method of "graph decomposition" is introduced. This method is essentially a technique of finding the network function of a network by a suitable topological decomposition dividing the original graph into several subgraphs whose network functions can be directly obtained from the tables. From the network functions of the subgraphs obtained first, the desired network function of the original network can be found.

II. Basic Graphs, Their Trees and Two-Trees

There is almost no science that can be developed without classification. Before going further, we will make a classification of linear graphs according to their nullities and their ranks. We call each graph a "basic graph". The trees of thirty-four basic graphs are given in Table 1. Any network function of a network whose nullity is less than or equal to six (if it is equal to six, the rank is no greater than six) can be obtained from the table shown in the following section.

Of all network functions, more commonly used ones are the driving-point immittance function, the transfer immittance function and the voltage-ratio transfer function. The topological expressions of these functions are[7]

$$z_{dr}^{ii'} = \frac{1}{y_{dr}^{ii'}} = \frac{V_{ii'}}{I_{ii'}} = \frac{\Sigma\, T_2^{ii'}}{\Sigma\, T}$$

$$z_{tr}^{ii'-jj'} = z_{tr}^{jj'-ii'} = \left.\frac{V_{jj'}}{I_{ii'}}\right|_{I_{jj'}=0} = \frac{\Sigma\, T_2^{ij-i'j'} - \Sigma\, T_2^{ij'-i'j}}{\Sigma\, T}$$

$$G^{ii'-jj'} = \frac{V_{jj'}}{V_{ii'}} = \frac{\frac{V_{jj'}}{I_{ii'}}}{\frac{V_{ii'}}{I_{ii'}}} = \frac{z_{tr}^{jj'-ii'}}{z_{dr}^{ii'}} = \frac{\Sigma\, T_2^{ij-i'j'} - \Sigma\, T_2^{ij'-i'j}}{\Sigma\, T_2^{ii'}}$$

where ii' and jj' denote terminal-pair nodes. $z_{dr}^{ii'}$ is the driving-point impedance. $y_{dr}^{ii'}$ is the driving-point admittance. $z_{tr}^{ii'-jj'}$ and $z_{tr}^{jj'-ii'}$ are transfer impedances. $\Sigma\, T$ is the summation of all tree admittance products. $\Sigma\, T_2^{ii'}$ is the summation of all two-tree admittance products, provided that nodes i and i' are in the different connected parts. $\Sigma\, T_2^{ij-i'j'}$ is the summation of the two-tree admittance products, provided that nodes i and j are in one connected part and the nodes i' and j' are in the other connected part, $\Sigma\, T_2^{ij'-i'j}$ is the summation of the two-tree admittance products, under the requirement that nodes i and j' are in one connected part and the nodes i' and j are in the other connected part.

Several instructions for finding the topological formulas by using the table are given below:

(1) Find the two-trees $T_2^{ii'}$ from the trees of a graph. By our definition, a two-tree is a pair of unconnected circuitless subgraphs, each subgraph being connected, which together include all the vertices of the graph. Therefore, choose the tree which contains the element $e_{ii'}$, the deletion of which will yield a two-tree. The symbol $t_2^{ii'}$ will be used for a two-tree. To find all the two-trees from the trees gives the total two-trees $T_2^{ii'}$.

(2) Find $T_2^{ij-i'j'}$ and $T_2^{ij'-i'j}$. To find $T_2^{ij-i'j'}$, we find the two trees of $T_2^{ii'}$ of which j and j' are in the separate connected parts, i and j are in one connected part, and i' and j' are in the other connected part. $T_2^{ij'-i'j}$ is obtained by the same way as $T_2^{ij-i'j'}$ except j and j' are interchanged.

(3) If between two nodes, there are more elements than one, take one at a time to combine with the other single elements. If this subgraph is one of the thirty-four basic graphs, by superposition the total trees would be the sum of the trees of all subgraphs.

We see that if the graph of a network is in the table, the trees and the two-trees of the graph can then be found. Since the topological formulas of network functions are in terms of only trees and two-trees, any network function of the graph can be obtained.

III. Find the Topological Formulas by Use of Table 1.

As shown in the previous section, if the graph of a network is exactly the same as one of the graphs given in the table, we find that the network function can be obtained directly by using the table. In the case that the graph of a network is not the same as any graph in the table, then we cannot use the table to find the network function at once. However, if the nullity of the graph is less than six (if the nullity of the graph is six, the rank is not greater than six), we can use a technique to make it be the one which is given in the table. Then the table will still be applicable for finding the network function of the graph. The nullities and ranks of the graphs of many commonly used networks in network analysis and network synthesis are usually within this range. This method is illustrated by the following two examples.

Example 1: Find the topological expressions of z_{dr}^{12} and z_{tr}^{12-56} of the LC ladder network in Fig. 1.

The graph of the network in Fig. 1(a) is shown in Fig. 1(b). We see that the graph is not contained in the table. However, we can add a fictitious element e_{25}. Now the new graph is G_2 in Table 1 whose trees are known. By doing exactly the same as if e_{25} existed, after z_{dr}^{12} and z_{tr}^{12-56} are obtained from N_2, either by setting $e_{25} = 0$ or by deleting the trees and two-trees which contain the fictitious element e_{25}, gives us z_{dr}^{12} and z_{tr}^{12-56} of the original graph. Following the procedure given above, we find the trees and the two-trees of the original graph are:

$$\begin{aligned}
\Sigma T = {} & y_{12}\, y_{13}\, y_{14}\, y_{15} + y_{12}\, y_{13}\, y_{14}\, y_{45} + y_{12}\, y_{13}\, y_{15}\, y_{34} + y_{12}\, y_{13}\, y_{15}\, y_{45} \\
& + y_{12}\, y_{13}\, y_{34}\, y_{45} + y_{12}\, y_{14}\, y_{15}\, y_{23} + y_{12}\, y_{14}\, y_{15}\, y_{34} + y_{12}\, y_{14}\, y_{23}\, y_{45} \\
& + y_{12}\, y_{14}\, y_{34}\, y_{45} + y_{12}\, y_{15}\, y_{23}\, y_{34} + y_{12}\, y_{15}\, y_{23}\, y_{45} + y_{12}\, y_{15}\, y_{34}\, y_{45} \\
& + y_{12}\, y_{23}\, y_{34}\, y_{45} + y_{13}\, y_{14}\, y_{15}\, y_{23} + y_{13}\, y_{14}\, y_{23}\, y_{45} + y_{13}\, y_{15}\, y_{23}\, y_{34} \\
& + y_{13}\, y_{15}\, y_{23}\, y_{45} + y_{13}\, y_{23}\, y_{34}\, y_{45} + y_{14}\, y_{15}\, y_{23}\, y_{34} + y_{14}\, y_{23}\, y_{34}\, y_{45} \\
& + y_{15}\, y_{23}\, y_{34}\, y_{45} \cdots\cdots\cdots\cdots \text{(A)}
\end{aligned}$$

$$\begin{aligned}
\Sigma T_2^{12} = {} & y_{13}\, y_{14}\, y_{15} + y_{13}\, y_{14}\, y_{45} + y_{13}\, y_{15}\, y_{34} + y_{13}\, y_{15}\, y_{45} \\
& + y_{13}\, y_{34}\, y_{45} + y_{14}\, y_{15}\, y_{23} + y_{14}\, y_{15}\, y_{34} + y_{14}\, y_{23}\, y_{45} \\
& + y_{14}\, y_{34}\, y_{45} + y_{15}\, y_{23}\, y_{34} + y_{15}\, y_{23}\, y_{45} + y_{15}\, y_{34}\, y_{45} \\
& + y_{23}\, y_{34}\, y_{45} \cdots\cdots\cdots\cdots \text{(B)}
\end{aligned}$$

$$\begin{aligned}
\Sigma T_2^{16-25} &= y_{23}\, y_{34}\, y_{45} \\
\Sigma T_2^{15-26} &= 0 \\
\Sigma T_2^{16-25} - \Sigma T_2^{15-26} &= y_{23}\, y_{34}\, y_{45} \cdots\cdots\cdots\cdots \text{(C)}
\end{aligned}$$

Therefore we obtain

$$z_{dr}^{12} = \frac{\Sigma T_2^{12}}{\Sigma T} = \frac{B}{A}$$

$$z_{tr}^{12-56} = \frac{\Sigma T_2^{16-25} - \Sigma T_2^{15-26}}{\Sigma T} = \frac{C}{A}$$

Example 2: Find the voltage-ratio transfer function G^{12-46} in topological terms of the RC Twin-T network in Fig. 2(a).

The network topology of the RC Twin-T network is shown in Fig. 2(b). We find that a Twin-T network can also be drawn as G_2 by adding two fictitious elements as shown in Fig. 2(c). By the same procedure as before, we get

$$\begin{aligned}\Sigma T = {} & y_{13}\, y_{15}\, y_{23}\, y_{35} + y_{13}\, y_{15}\, y_{23}\, y_{45} + y_{13}\, y_{15}\, y_{34}\, y_{52} \\ & + y_{13}\, y_{15}\, y_{45}\, y_{52} + y_{13}\, y_{23}\, y_{34}\, y_{45} + y_{13}\, y_{23}\, y_{34}\, y_{52} \\ & + y_{13}\, y_{23}\, y_{45}\, y_{52} + y_{13}\, y_{34}\, y_{45}\, y_{52} + y_{15}\, y_{23}\, y_{34}\, y_{45} \\ & + y_{15}\, y_{23}\, y_{34}\, y_{52} + y_{15}\, y_{23}\, y_{45}\, y_{52} + y_{15}\, y_{34}\, y_{45}\, y_{52} . \quad . \; . \; . \; (D)\end{aligned}$$

$$\begin{aligned}\Sigma T_2^{12} = {} & y_{13}\, y_{15}\, y_{34} + y_{13}\, y_{15}\, y_{45} + y_{13}\, y_{34}\, y_{45} + y_{13}\, y_{34}\, y_{52} \\ & + y_{13}\, y_{45}\, y_{52} + y_{15}\, y_{23}\, y_{34} + y_{15}\, y_{23}\, y_{45} + y_{15}\, y_{34}\, y_{45} \\ & + y_{23}\, y_{34}\, y_{45} + y_{23}\, y_{34}\, y_{52} + y_{23}\, y_{45}\, y_{52} + y_{34}\, y_{45}\, y_{52} \quad . \; . \; . \; (E)\end{aligned}$$

$$\Sigma T_2^{16-24} = y_{23}\, y_{34}\, y_{45} + y_{23}\, y_{34}\, y_{25} + y_{25}\, y_{45}\, y_{34} + y_{25}\, y_{45}\, y_{23}$$

$$\Sigma T_2^{14-26} = 0$$

$$\Sigma T_2^{16-24} - \Sigma T_2^{14-26} = y_{23}\, y_{34}\, y_{45} + y_{23}\, y_{34}\, y_{25} + y_{25}\, y_{45}\, y_{34} + y_{25}\, y_{45}\, y_{23}$$

$$. \; . \; . \; . \; . \; . \; . \; . \; . \; . \; (F)$$

Thus,

$$z_{dr}^{12} = \frac{\Sigma T_2^{12}}{\Sigma T} = \frac{E}{D}$$

$$z_{tr}^{12-46} = \frac{\Sigma T_2^{16-24} - \Sigma T_2^{14-26}}{\Sigma T} = \frac{F}{D}$$

$$G^{12-46} = \frac{V_{46}}{V_{12}} = \frac{z_{tr}^{12-46}}{z_{dr}^{12}} = \frac{\frac{E}{D}}{\frac{F}{D}} = \frac{E}{F}$$

Two examples shown above were dealing with two low nullity graphs. The higher the nullity and rank a graph is, the more trees and two-trees it contains. If we used a higher nullity and rank in the example, it would take more space to write the trees and two-trees. This is needless since the method is as valid for higher nullity and rank graphs as for lower ones.

IV. A Method of Finding Topological Formulas of an n-node Graph from its (n-1)-node Subgraph

In this section, the graphs that are dealt with are the graphs whose nullities and ranks are beyond those of the graphs given in the table. For those higher nullity and rank graphs, the network function cannot be obtained by using the previous method. A method given below shows that the network function of a high nullity and rank graph can be easily obtained by use of the table, and a lot of labor will be saved. This method is in fact a method of finding all the trees of an n-node graph from the trees of (n-1)-node subgraph. After the trees of the given graph are obtained, using the method as given in Sec. II, the network function can be found.

Consider the general n-node graph in Fig. 3. Note that a dotted-line circle is cutting all the elements that are connecting the node n_n with the nodes $n_1, n_2, \ldots, n_{n-1}$. If one takes out these elements, the remaining graph in the circle is a subgraph of the original graph containing (n-1) nodes.

Definition 1: Bisecting circle - A bisecting circle is a circle drawn on a connected graph dividing the graph into two parts. We shall call the subgraph that is inside the circle an interior subgraph, and the subgraph outside, exterior subgraph. The elements that are cut by the circle belong to neither of the two. A bisecting circle is not allowed to have any node on it.

Definition 2: Interior node and exterior node - A node which is inside a bisecting circle is termed an interior node with respect to this bisecting circle; one outside the bisecting circle, an exterior node.

Definition 3: Interior element and exterior element - An element connecting between two interior nodes is an interior element. An element connecting between an exterior node and an interior node, or connecting between two exterior nodes is an exterior element.

Definition 4: Ex-incident node - The interior node which is incident with an exterior element or more is an ex-incident node. The interior node which is not incident with any exterior element is a non-ex-incident node.

Definition 5: Ex-incident element - The interior element connecting two ex-incident nodes is an ex-incident element. The interior element which is not an ex-incident element is a non-ex-incident element.

Definition 6: Ex-incident subgraph - A subgraph of a graph which contains one or more ex-incident elements is an ex-incident subgraph. A subgraph of a graph which contains no ex-incident elements is called a non-ex-incident subgraph.

Definition 7: Interior tree - A tree of an interior subgraph is an interior tree.

Definition 8: Order and degree - The number of exterior elements contained in a graph, subgraph or tree is the order of that graph, subgraph or tree. The number of ex-incident elements contained in a graph, subgraph or tree is the degree of that graph, subgraph or tree.

In Fig. 3, the n-node graph is divided by a bisecting circle into two parts. The interior subgraph has (n-1) nodes while the exterior subgraph has only one. For this case, we have the following theorem:

Theorem: If a connected n-node graph is cut by a bisecting circle, and the exterior subgraph contains only a single node, (a) a interior tree plus any exterior element is a tree of the graph, and (b) the total trees is the summation of the first, second, ... , nth order trees.

Proof: (a) An interior tree is a connected graph containing (n-1) nodes and (n-2) elements. After an exterior element is added, the new subgraph will have n nodes and (n-1) elements. It is still connected. By definition it is a tree of the graph. (b) By definition, a tree is a connected graph containing all the nodes, thus every tree contains at least one exterior element. Grouping the trees according to its order yields the first, second, ... , nth order trees. The union of these sets is the total trees.

The procedure of finding the ith order trees is given as follows:

(1) Select i exterior elements at a time and add them to each interior tree.

(2) Pick out all the ex-incident subgraphs and delete (i-1) interior

elements which must include all the ex-incident elements*. This will yield ith order trees.

(3) Also, delete (i-1) interior elements from each non-ex-incident subgraph. Select all the ith order trees that were not obtained in (2).

(4) Repeat the process for all the combinations of i exterior elements. The sum of all the ith order trees obtained from the ex-incident subgraphs and the non-ex-incident subgraphs of all the combinations of i exterior elements gives the total ith order trees.

The method of finding all the trees of an n-node graph from its (n-1)-node subgraph therefore simply contains two steps:

(1) Find the first, second, ... , nth order trees from the trees of the (n-1)-node subgraph;

(2) Summarize all these sets of trees giving the total trees.

As stated above, this method is essentially used in the case when the nullity and rank of a graph is beyond those of the graphs given in the table.

Because of the graphs of high nullity and rank containing a lot of trees, it will take a large space to write them down. In interest of size, N_2 is again to be used as an example.

Example 3: Find the trees of N_2 using the method of finding the trees of an n-node graph with the help of the trees of its (n-1)-node subgraph.

In Fig. 4, the interior trees of Fig. 4(a) is shown in Fig. 4(b).

(1) Find first, second, third order trees.

(i) First order trees - The first order trees are obtained directly from adding e_{25}, e_{12} and e_{23} one at a time to each interior tree. Thus the number of total first order trees is 24.

(ii) Second order trees - First we shall find the second order trees of e_{25} and e_{12}, and e_{12} and e_{23}. After e_{25} are added to each interior tree, the ex-incident subgraphs are found as shown in Fig. 4(c).

---

*An ith order graph, whose interior subgraph is a tree, can at most have (i-1) ex-incident elements.

Deleting the ex-incident element from the ex-incident subgraphs gives the second order trees of e_{25} and e_{12} as shown in Fig. 4(d). By checking with the second order trees obtained from the non-ex-incident subgraphs, we find no other second order tree which is not included in these five. Hence the total second order trees of e_{25} and e_{12} is five. Because of symmetry, the second order trees of e_{12} and e_{23} can similarly be found. The sum of these two sets of second order trees is ten. The second order trees of e_{25} and e_{23} as shown in Fig. 4(e) are found from the non-ex-incident subgraphs. Therefore the number of total second order trees is eighteen.

(iii) Third order trees - After the three exterior elements are added to each interior tree as shown in Fig. 4(f), three third order trees, obtained by deleting the two ex-incident subgraphs, are shown in Fig. 4(g). By checking, we find there is no other third order tree other than these trees. The total third order tree is three.

(2) Summing the first, second and third order trees, 24 + 18 + 3 = 45, we get the total forty-five trees.

For high nullity and rank graphs, we may use the technique of cutting the graph by several bisecting circles as shown in Fig. 5. Use the method once for each bisecting circle from inner to outer, step by step, the trees of the total graph can therefore be found.

V. Graph Decomposition Method

The method introduced in the previous section is a quite useful method for finding the network function of a network whose nullity is not very much higher than six. As was mentioned before, in determining the network function for high-nullity and high-rank graphs, this method requires several bisecting circles. The trees of the total graph can be obtained by applying the method step by step, once for each bisecting circle from the inner circle to the outer circle. Apparently, this method is a tedious task for the high-nullity and high-rank graphs. Another method is given for finding the network function particularly for high-nullity and high-rank networks also by using the table, namely, the graph decomposition method.

Consider the symbolic diagram in Fig. 6. Suppose the circle G is a finite graph of high-nullity and high-rank. The graph decomposition method is essentially a method of finding the network function by decompositing the original graph into subgraphs in such a way that:

(1) No element can be split

(2) The junctions between the subgraphs are only nodes

(3) The network function of each subgraph can be obtained by using the previous methods

(4) The connection between the subgraphs must be one of the connections as shown in Fig. 7, or a combination of them.

(A) One port

(a) Series connection. The driving-point impedance is

$$Z_{dr} = Z_{dr}^{(1)} + Z_{dr}^{(2)} + \dots + Z_{dr}^{(n)}$$

(b) Parallel connection. The driving-point admittance is

$$Y_{dr} = Y_{dr}^{(1)} + Y_{dr}^{(2)} + \dots + Y_{dr}^{(n)}$$

(B) Two port

(a) Series connection. The impedance matrix is

$$[Z] = [Z^{(1)}] + [Z^{(2)}] + \dots + [Z^{(n)}]$$

(b) Parallel connection. The admittance matrix is

$$[Y] = [Y^{(1)}] + [Y^{(2)}] + \dots + [Y^{(n)}]$$

(c) Cascade connection. The transmission matrix is

$$\begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{bmatrix} = \begin{bmatrix} \mathcal{A}^{(1)} & \mathcal{B}^{(1)} \\ \mathcal{C}^{(1)} & \mathcal{D}^{(1)} \end{bmatrix} \begin{bmatrix} \mathcal{A}^{(2)} & \mathcal{B}^{(2)} \\ \mathcal{C}^{(2)} & \mathcal{D}^{(2)} \end{bmatrix} \cdots \begin{bmatrix} \mathcal{A}^{(n)} & \mathcal{B}^{(n)} \\ \mathcal{C}^{(n)} & \mathcal{D}^{(n)} \end{bmatrix}$$

A typical combinated connection is shown in Fig. 8.

Example 4: It is required to find the transmission matrix of the network shown in Fig. 9(a).

It is seen that the graph of the network in Fig. 9(b) can be decomposited into two parts which are topologically identical. In Example 2, it was found that for G_1 the driving-point impedance and transfer impedance are

$$Z_{dr}^{12} = \frac{\Sigma T_2^{12}}{\Sigma T} = \frac{E}{D}$$

$$Z_{tr}^{12-46} = Z_{tr}^{46-12} = \frac{\Sigma T_2^{16-24} - \Sigma T_2^{14-26}}{\Sigma T} = \frac{F}{D}$$

Since G_1 is symmetric, ΣT_2^{26} can be directly obtained by changing the subscript 2 to 4 and the subscript 4 to 2 in the equation (E) which is

$$\left.\Sigma T_2^{46}\right|_{G_1} = y_{13} y_{15} y_{32} + y_{13} y_{15} y_{25} + y_{13} y_{32} y_{25} + y_{13} y_{32} y_{54}$$

$$+ y_{13} y_{25} y_{54} + y_{15} y_{43} y_{32} + y_{15} y_{43} y_{25} + y_{15} y_{32} y_{25}$$

$$+ y_{43} y_{32} y_{25} + y_{43} y_{32} y_{54} + y_{43} y_{25} y_{54} + y_{32} y_{25} y_{54} \quad \text{.......(G)}$$

then

$$\left.Z_{dr}^{46}\right|_{G_1} = \frac{G}{D}$$

For G_2, let $H = \left.\Sigma T\right|_{G_2}$, $I = \left.\Sigma T_2^{1'2'}\right|_{G_2}$, $J = \left.\Sigma T_2^{1'6-2'4}\right|_{G_2} - \left.\Sigma T_2^{1'4-2'6}\right|_{G_2}$

and $K = \left.\Sigma T_2^{46'}\right|_{G_2}$ Similarly, H, I, J and K can be found by changing the subscripts 2, 3, 5 to 2', 3', 5', respectively in the equations (D), (E), (F) and (G). We find that for G_2

$$Z_{dr}^{1'2'} = \frac{I}{H}$$

$$Z_{tr}^{1'2'-46} = Z_{tr}^{46-1'2'} = \frac{J}{H}$$

$$Z_{dr}^{46}\Big|_{G_2} = \frac{K}{H}$$

Note that in the tandem connection the output voltage of the first two-port is equal to the input voltage of the second two-port, whereas the output current of the first two-port is the negative of the input current of the second two-port

$$V_2 = V_1'$$

$$-I_2 = I_1'$$

Therefore

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} \mathcal{A}^{(1)} & \mathcal{B}^{(1)} \\ \mathcal{C}^{(1)} & \mathcal{D}^{(1)} \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix}$$

$$\begin{bmatrix} V_1' \\ I_1' \end{bmatrix} = \begin{bmatrix} \mathcal{A}^{(2)} & \mathcal{B}^{(2)} \\ \mathcal{C}^{(2)} & \mathcal{D}^{(2)} \end{bmatrix} \begin{bmatrix} V_2' \\ -I_2' \end{bmatrix}$$

where

$$\mathcal{A}^{(1)} = \frac{Z_{dr}^{12}}{Z_{tr}^{12-46}} = \frac{\frac{E}{D}}{\frac{F}{d}} = \frac{E}{F}$$

$$\mathcal{B}^{(1)} = \frac{Z_{dr}^{12}\, Z_{dr}^{46}\Big|_{G_1} - (Z_{tr}^{12-46})^2}{Z_{tr}^{12-46}} = \frac{\frac{E}{D} \cdot \frac{G}{D} - \frac{F^2}{D^2}}{\frac{F}{D}} = \frac{EG - F^2}{FD}$$

$$\mathcal{C}^{(1)} = \frac{1}{Z_{tr}^{12-46}} = \frac{1}{\frac{F}{D}} = \frac{D}{F}$$

$$\mathcal{B}^{(1)} = \frac{\left. Z_{dr}^{46} \right|_{G_1}}{Z_{tr}^{12-46}} = \frac{\frac{G}{D}}{\frac{F}{D}} = \frac{G}{F}$$

$$\mathcal{A}^{(2)} = \frac{\left. Z_{dr}^{46} \right|_{G_2}}{Z_{tr}^{46-1'2'}} = \frac{\frac{K}{H}}{\frac{J}{H}} = \frac{K}{J}$$

$$\mathcal{B}^{(2)} = \frac{\left. Z_{dr}^{46} \right|_{G_2} Z_{dr}^{1'2'} - (Z_{tr}^{46-1'2'})^2}{Z_{tr}^{46-1'2'}} = \frac{\frac{K}{H} \cdot \frac{I}{H} - (\frac{J}{H})^2}{\frac{J}{H}} = \frac{KI - J^2}{JH}$$

$$\mathcal{C}^{(2)} = \frac{1}{Z_{tr}^{46-1'2'}} = \frac{1}{\frac{J}{H}} = \frac{H}{J}$$

$$\mathcal{D}^{(2)} = \frac{Z_{dr}^{1'2'}}{Z_{tr}^{46-1'2'}} = \frac{\frac{I}{H}}{\frac{J}{H}} = \frac{I}{J}$$

The transmission matrix of G is

$$\begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{bmatrix} = \begin{bmatrix} \mathcal{A}^{(1)} & \mathcal{B}^{(1)} \\ \mathcal{C}^{(1)} & \mathcal{D}^{(1)} \end{bmatrix} \begin{bmatrix} \mathcal{A}^{(2)} & \mathcal{B}^{(2)} \\ \mathcal{C}^{(2)} & \mathcal{D}^{(2)} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{D\,E\,K + H(E\,G - F^2)}{F\,D\,J} & \dfrac{D\,E(K\,I - J^2) + I\,H(E\,G - F^2)}{D\,F\,H\,J} \\ \dfrac{D\,K + G\,I}{F\,J} & \dfrac{D(K\,I - J^2) + H\,G\,I}{F\,J\,H} \end{bmatrix}$$

Example 5: It is desired to find the Z matrix of the specified two-port graph as shown in Fig. 10(a).

By inspection, it is seen that this graph can be decomposited into two subgraphs in parallel as shown in Fig. 10(b). This is a graph corresponding to two two-port ladder networks properly connected in parallel. The driving-point impedance and the transfer impedance of G_1 were found in Example 1 as:

$$Z_{dr}^{12} = \Sigma T_2^{12} = \frac{B}{A}$$

$$Z_{tr}^{12-15} = Z_{tr}^{15-12} = \frac{\Sigma T_2^{11-25} - \Sigma T_2^{15-21}}{\Sigma T} = \frac{C}{A}$$

ΣT_2^{15} can be obtained by replacing the subscript 2 by 5 and the subscript 5 by 2, that is

$$\begin{aligned} \Sigma T_2^{15} = {} & y_{13} y_{14} y_{12} + y_{13} y_{14} y_{42} + y_{13} y_{12} y_{34} + y_{13} y_{12} y_{42} \\ & + y_{13} y_{34} y_{42} + y_{14} y_{12} y_{53} + y_{14} y_{12} y_{34} + y_{14} y_{53} y_{42} \\ & + y_{14} y_{34} y_{42} + y_{12} y_{53} y_{34} + y_{12} y_{53} y_{42} + y_{12} y_{34} y_{42} \\ & + y_{53} y_{34} y_{42} \quad \text{. . . . .} \quad \text{(L)} \end{aligned}$$

Thus

$$Z_{dr}^{15} = \frac{\Sigma T_2^{15}}{\Sigma T} = \frac{L}{A}$$

The impedance matrix of G_1 is

$$[Z^{(1)}] = \left[\begin{matrix} Z_{dr}^{12} & Z_{tr}^{12-15} \\ Z_{tr}^{15-12} & Z_{dr}^{15} \end{matrix}\right]_{G_1} = \left[\begin{matrix} \frac{B}{A} & \frac{C}{A} \\ \frac{C}{A} & \frac{L}{A} \end{matrix}\right]$$

If we let $M = \Sigma T$, $N = \Sigma T_2^{12}$, $P = \Sigma T_2^{11-25} - \Sigma T_2^{15-21}$ and $Q = \Sigma T_2^{15}$ of G_2, M, N, P and Q can be obtained by replacing the subscripts 3 and 4 by 3' and 4', respectively in the equations (A), (B), (C) and (L). Then the impedance matrix of G_2 is

$$[Z^{(2)}] = \left[\begin{matrix} Z_{dr}^{12} & Z_{tr}^{12-15} \\ Z_{tr}^{15-12} & Z_{dr}^{15} \end{matrix}\right]_{G_2} = \left[\begin{matrix} \frac{N}{M} & \frac{P}{M} \\ \frac{P}{M} & \frac{Q}{M} \end{matrix}\right]$$

The over-all impedance matrix is

$$[Z] = [Z^{(1)}] + [Z^{(2)}] = \begin{bmatrix} \frac{B}{A} + \frac{N}{M} & \frac{C}{A} + \frac{P}{M} \\ \frac{C}{A} + \frac{P}{M} & \frac{L}{A} + \frac{Q}{M} \end{bmatrix}$$

VI. Conclusion

A new procedure for finding network functions has been given which contains the following two steps:

(1) Find the network topology, or graph, of the given network

(2) Determine the nullity from either the network or its graph,

(i) if it is less than six (if it is six, the rank is not greater that six), the network function can be found either directly from the table or by using the table after adding fictitious elements,

(ii) if it is not very much greater than six, the method of finding network functions of an n-node graph from its (n-1)-node subgraph is suggested to be used.

(iii) If the nullity and rank are high, then use the method of graph decomposition.

Consequently, by using the table and these methods, any network function of a finite graph can be obtained easily.

ACKNOWLEDGEMENT

The author is deeply indebted to Professor M. E. Van Valkenburg of the Department of Electrical Engineering at the University of Illinois for his valuable advice throughout the preparation of the manuscript. The author also wishes to express his gratitude to Mr. S. Y. Hwang for his help in various ways.

REFERENCES

1. Foster, R. M., "Topologic and Algebraic Considerations in Network Synthesis", Proc. Polytech. Inst. Brooklyn Symposium on Modern Network Synthesis I, 8-18 (April, 1952).

2. Seshu, S., "Topological Considerations in the Design of Driving Point Functions", Trans. IRE, Vol. CT-2, pp. 356-367, (December, 1955).

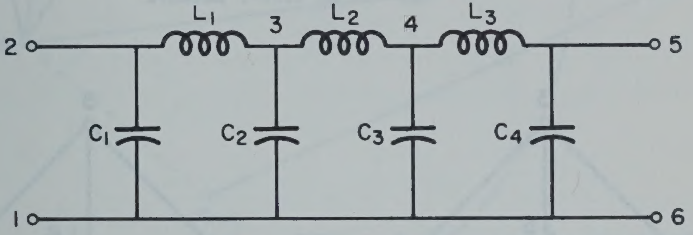
3. Mayeda, W. and Seshu, S., "Topological Formulas for Network Functions", Bulletin No. 446, U. of Ill. Engineering Experiment Station, (1957).

4. Van Valkenburg, M. E., "Topological Synthesis", IRE Wescon Record (Part 2) 2, 3-9, (1958).

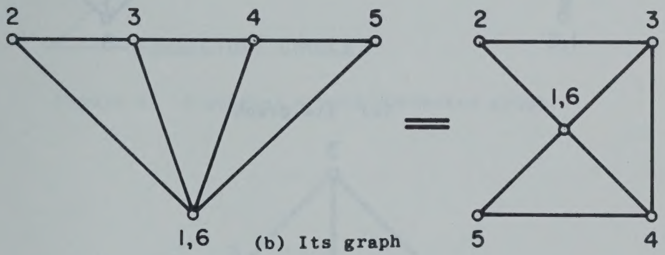
5. Kim, W. H., "Role of Network Topology in Network Synthesis", Proc. Midwest Symposium on Circuit Theory 3, 6.1-6.18, (1958).

6. Harary, F., "Graph Theory and Electric Networks", Trans. IRE, Vol. CT-6 (Special Supplement), 95-109, (May, 1959).

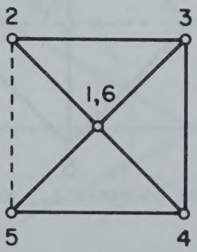
7. Mayeda, W., and Van Valkenburg, M. E., "Network Analysis and Synthesis by Digital Computer", IRE Conventional Record, (Part 2) 5, 137-144, (1957).



(a) An LC ladder network

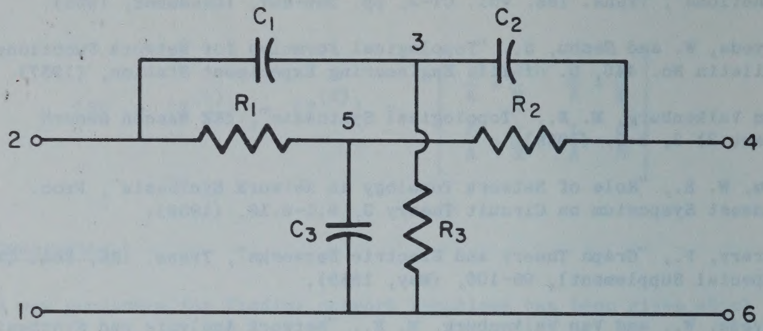


(b) Its graph

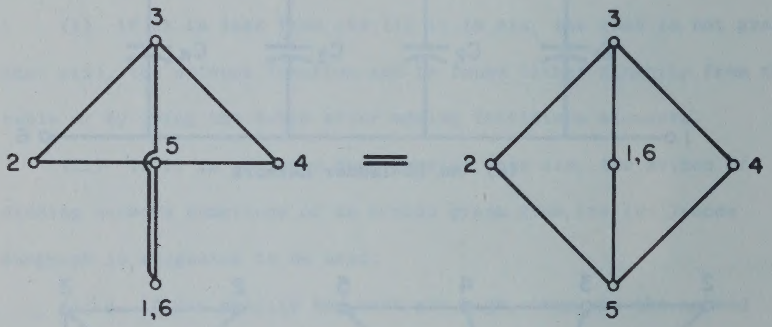


(c) Add a fictitious element e_{25} to the graph in (b) to form N_2 in Table 1.

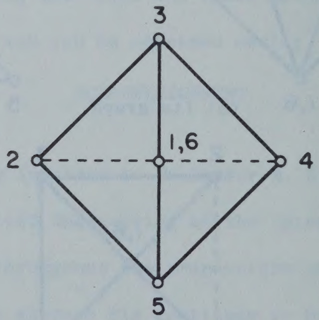
Figure 1. The network and its graph of example 1.


C_1
C_2
3
R_1
R_2
2
5
4
C_3
R_3
1
6

(a) An RC twin-T network


3
5
2
4
1,6
=
3
1,6
2
4
5

(b) Its graph



(c) Add two fictitious elements e_{21} and e_{41}, the graph becomes N_2.

Figure 2. The network and its graph of example 2.

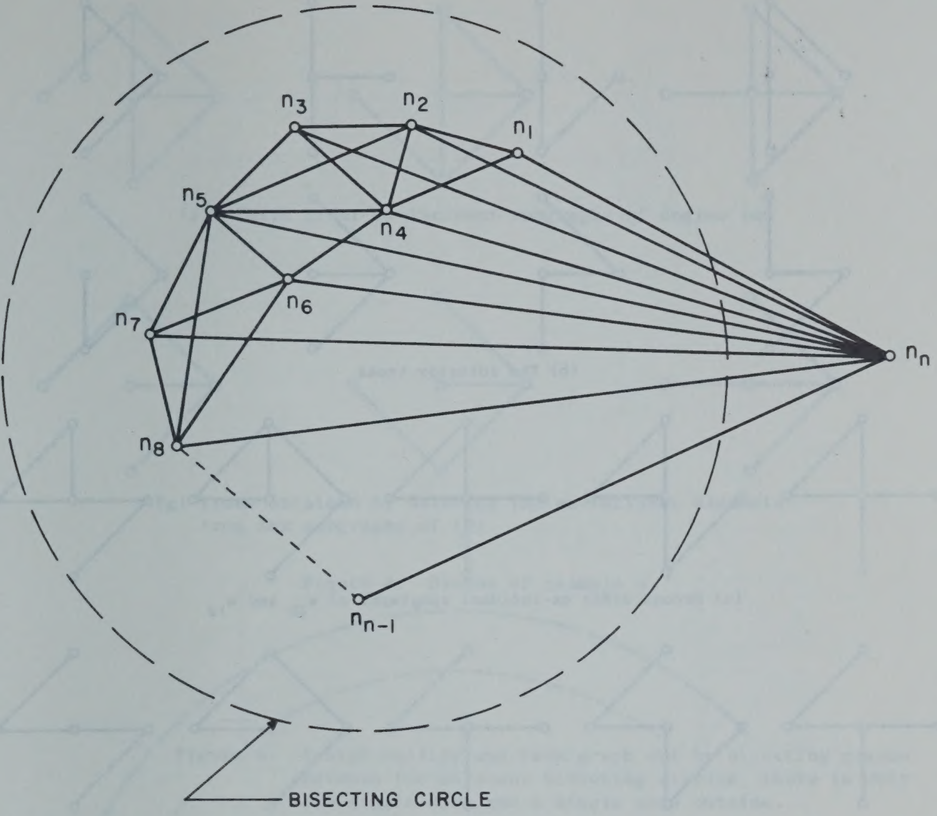
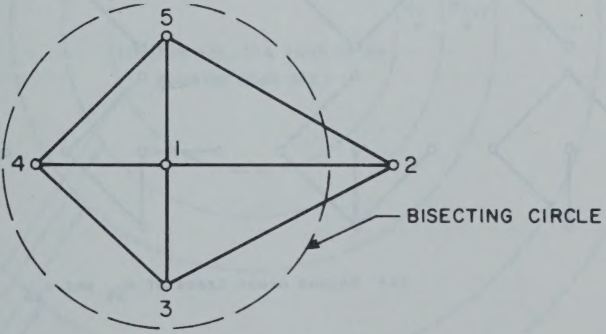
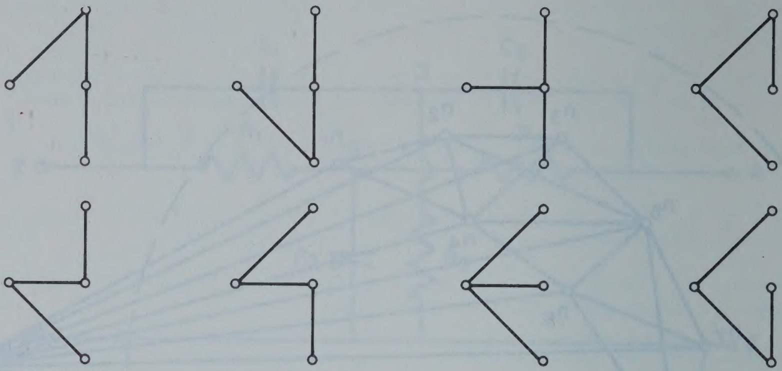

n₃
n₂
n₁
n₅
n₄
n₆
n₇
n₈
n_n
n_{n-1}
BISECTING CIRCLE

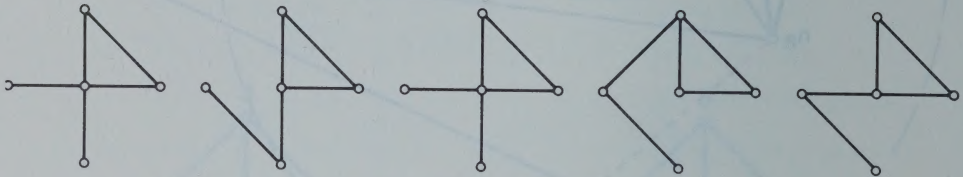
Figure 3. A general n-node connected graph.


5
4
1
2
3
BISECTING CIRCLE

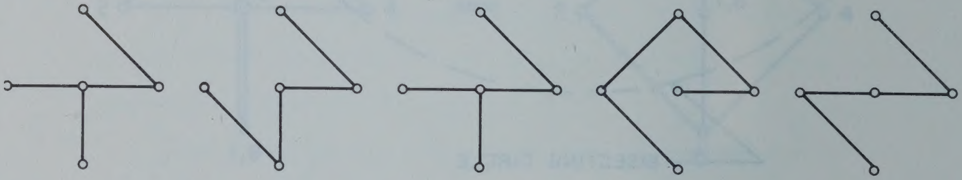
(a) N_2 cut by a bisecting circle



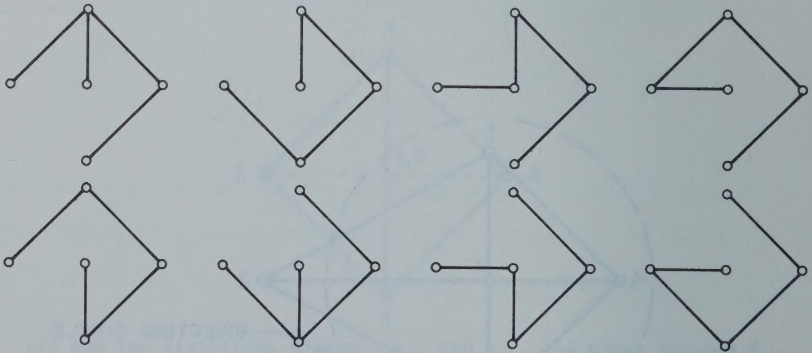
(b) The interior trees



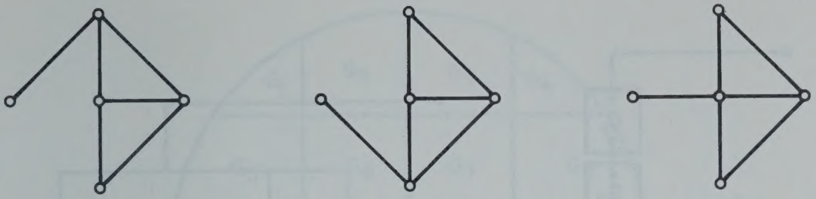
(c) Second order ex-incident subgraphs of e_{25} and e_{12}



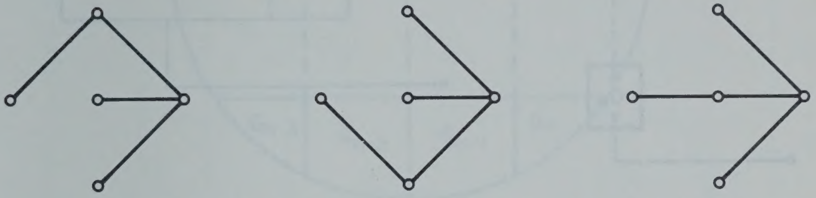
(d) Second order trees of e_{25} and e_{12}



(e) Second order trees of e_{25} and e_{23}



(f) Third order ex-incident subgraphs of degree two



(g) Trees obtained by deleting the ex-incident elements from the subgraphs of (f)

Figure 4. Graphs of example 3.

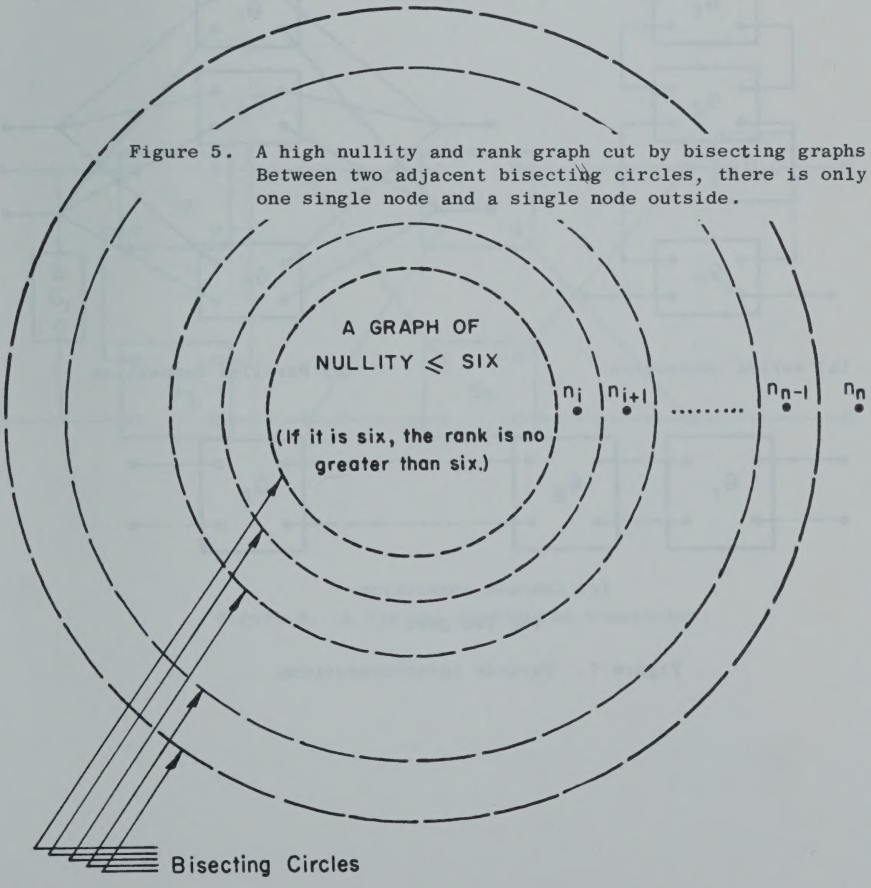
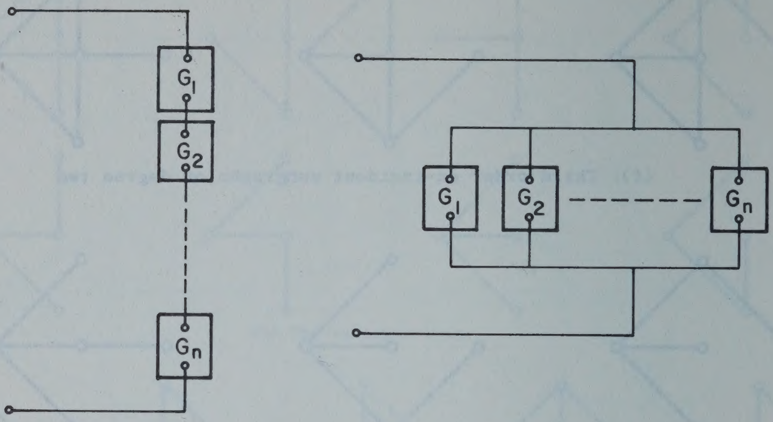
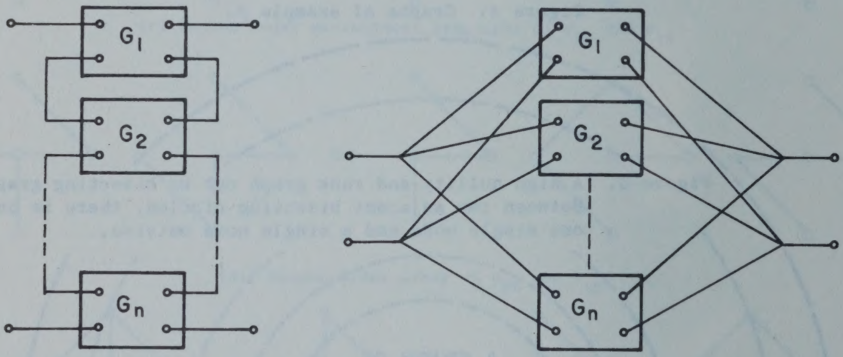

A GRAPH OF
NULLITY ≤ SIX
(If it is six, the rank is no greater than six.)
n_i
n_{i+1}
....
n_{n-1}
n_n
Bisecting Circles

Figure 5. A high nullity and rank graph cut by bisecting graphs. Between two adjacent bisecting circles, there is only one single node and a single node outside.

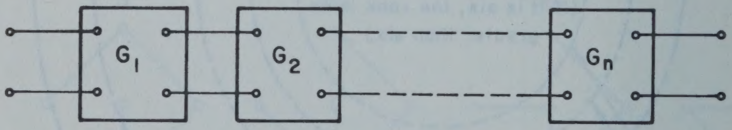

G1
G2
Gn
G1
G2
Gn

(a) Series connection (b) Parallel connection

(A) One port


G1
G2
Gn
G1
G2
Gn

(a) Series connection (b) Parallel connection



(c) Cascade connection

(B) Two port

Figure 7. Various interconnections

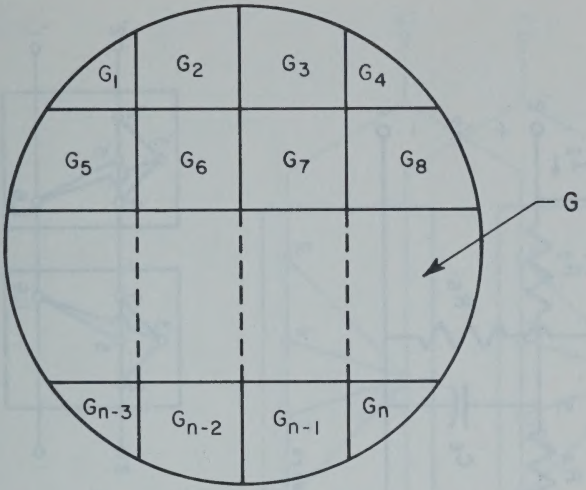

G_1
G_2
G_3
G_4
G_5
G_6
G_7
G_8
G
G_{n-3}
G_{n-2}
G_{n-1}
G_n

Figure 6. A symbolic representation of a finite graph.

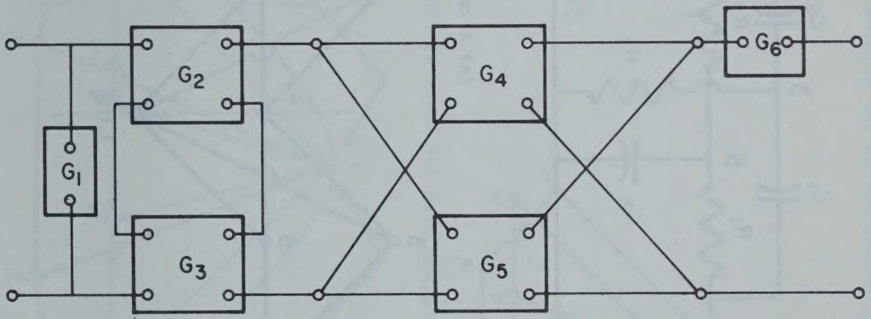
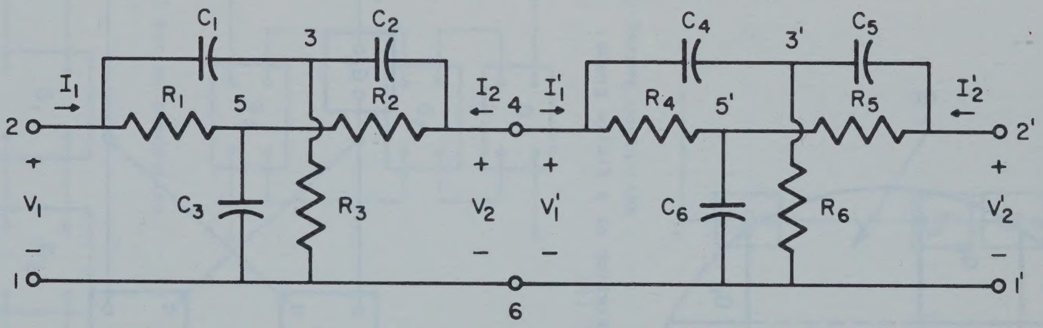
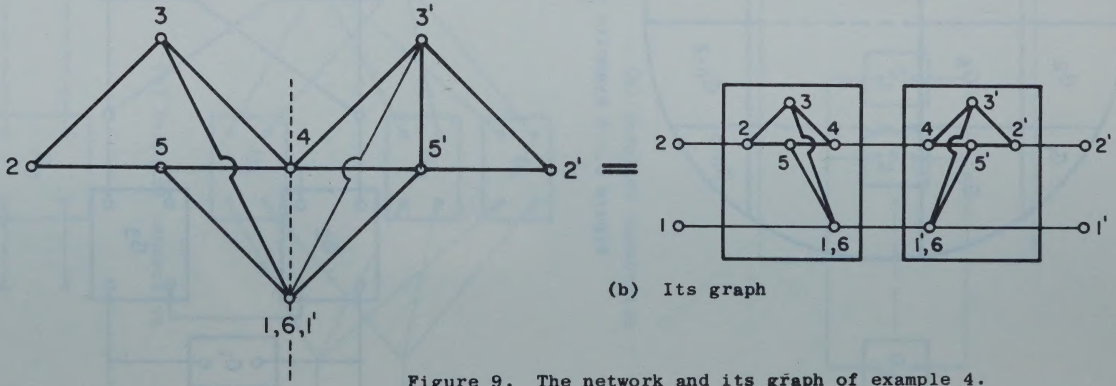

G_1
G_2
G_3
G_4
G_5
G_6

Figure 8. A typical combinated connection.

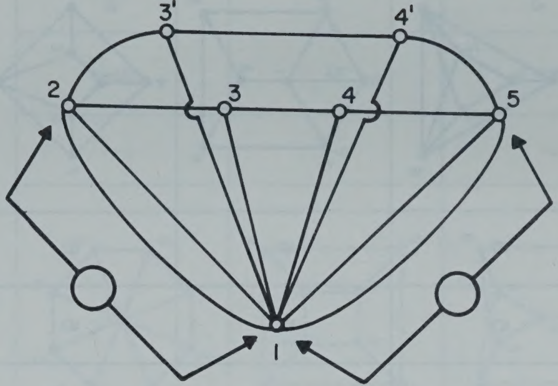

C1
3
C2
C4
3'
C5
I1
R1
5
R2
I2
4
I'1
R4
5'
R5
I'2
2
2'
+
V1
−
C3
R3
+
V2
−
+
V'1
−
C6
R6
+
V'2
−
1
6
1'

(a) A cascaded RC twin-T network

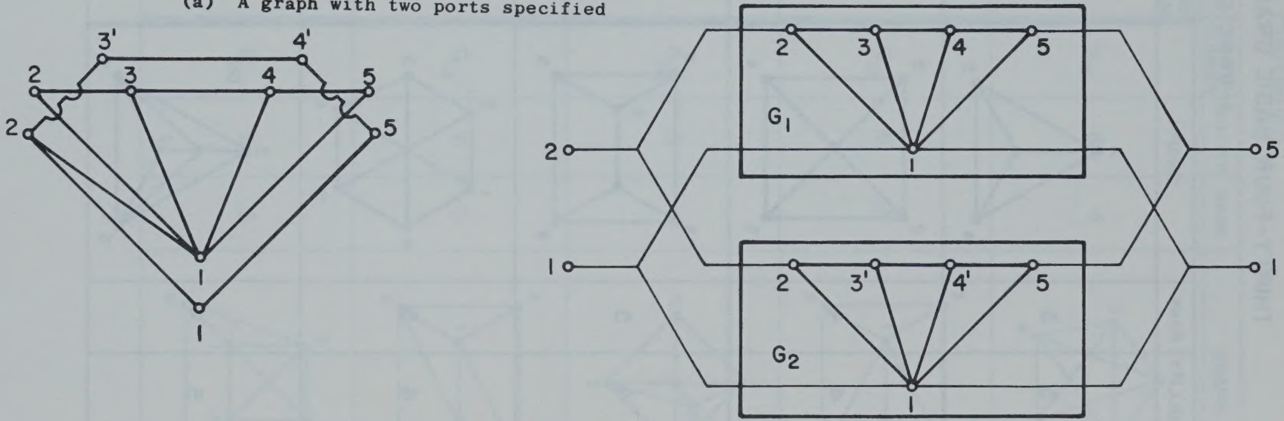

3
3'
2
5
4
5'
2'
1,6,1'
=
2
3
4
5
1,6
3'
4
2'
5'
1',6
2
1
2'
1'

(b) Its graph

Figure 9. The network and its graph of example 4.



(a) A graph with two ports specified


3'
4'
2
3
4
5
2
5
1
1
2
3
4
5
G1
1
2
5
1
1
2
3'
4'
5
G2
1

(b) Decompositing the original graph into two subgraphs in parallel

Figure 10. The graph and its decomposition of example 5.

TABLE 1.

THIRTY-FOUR BASIC GRAPHS AND THEIR TREES

A-BASIC GRAPHS

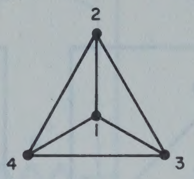
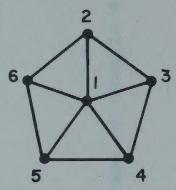
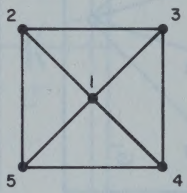
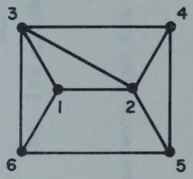
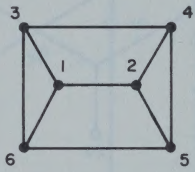
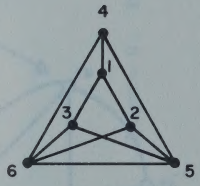
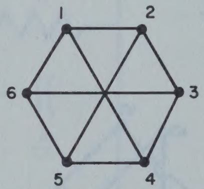
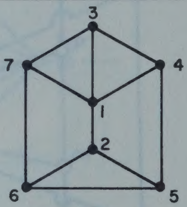
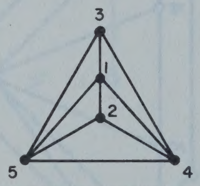
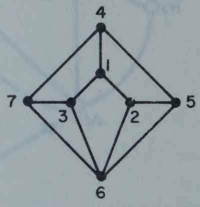
GRAPH NUMBER	NULLITY	RANK	GRAPH	GRAPH NUMBER	NULLITY	RANK	GRAPH
G_1	3	3		G_6	5	5	
G_2	4	4		G_7	5	5	
G_3	4	5		G_8	5	5	
G_4	4	5		G_9	5	6	
G_5	5	4		G_{10}	5	6	

TABLE 1. CONTINUED

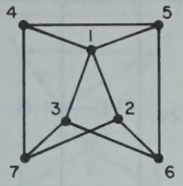
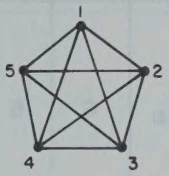
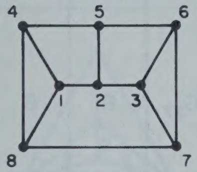
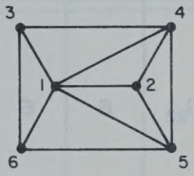
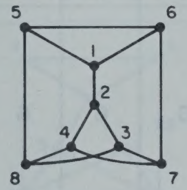
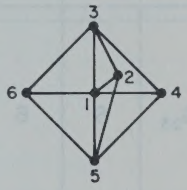
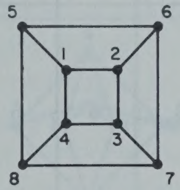
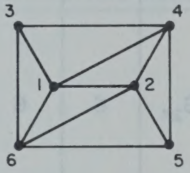
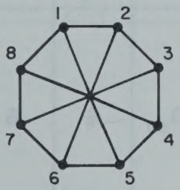
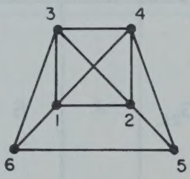
GRAPH NUMBER	NULLITY	RANK	GRAPH	GRAPH NUMBER	NULLITY	RANK	GRAPH
G_{11}	5	6		G_{16}	6	4	
G_{12}	5	7		G_{17}	6	5	
G_{13}	5	7		G_{18}	6	5	
G_{14}	5	7		G_{19}	6	5	
G_{15}	5	7		G_{20}	6	5	

TABLE 1. CONTINUED

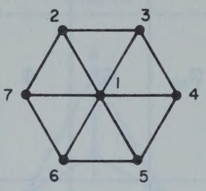
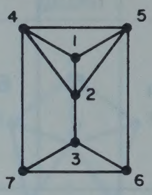
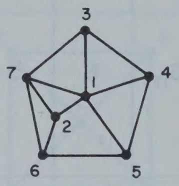
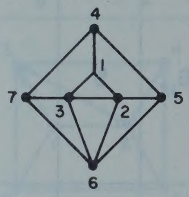
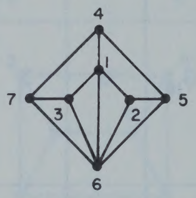
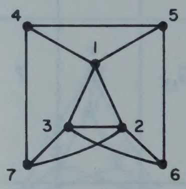
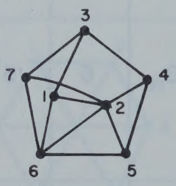
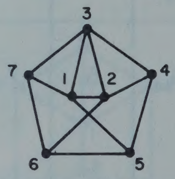
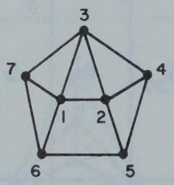
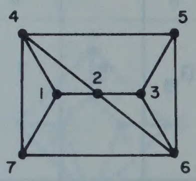
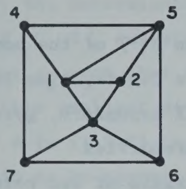
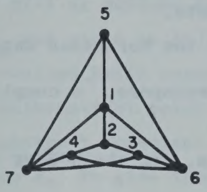
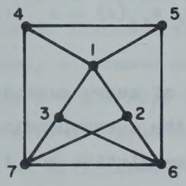
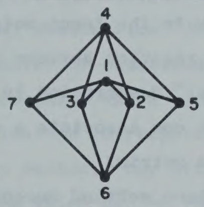
GRAPH NUMBER	NULLITY	RANK	GRAPH	GRAPH NUMBER	NULLITY	RANK	GRAPH
G_{21}	6	6		G_{26}	6	6	
G_{22}	6	6		G_{27}	6	6	
G_{23}	6	6		G_{28}	6	6	
G_{24}	6	6		G_{29}	6	6	
G_{25}	6	6		G_{30}	6	6	

TABLE 1. CONTINUED

GRAPH NUMBER	NULLITY	RANK	GRAPH	GRAPH NUMBER	NULLITY	RANK	GRAPH
G_{31}	6	6					
G_{32}	6	6					
G_{33}	6	6					
G_{34}	6	6					

ON A NEW CHARACTERIZATION OF LINEAR PASSIVE SYSTEMS*

R. E. KALMAN
Center for Control Theory
Research Institute for Advanced Studies (RIAS)
Baltimore 12, Maryland

1. Dynamical Equations of an N-Port.

Consider an N × N matrix function $Z(\cdot)$ of the complex variable s. Assume that it is the impedance matrix of a finite, time-invariant, passive RLCTT (resistor, inductor, capacitor, ideal transform, gyrator) N-port. Then, as is well known, $Z(\cdot)$ has the following properties

(i) Every element of $Z(\cdot)$ is a ratio of two relatively prime polynomials with real coefficients.

(ii) For every s with Re $s \geqq 0$, the hermetian matrix

$$Z(s) + Z'(\bar{s}) \quad (\ ' = \text{transpose}, \ ^{-} = \text{complex conjugate}) \tag{1.1}$$

is nonnegative definite.

If gyrators are not allowed, i.e., if the N-port is reciprocal, then we must have also

(iii) $Z(\cdot)$ is symmetrical, i.e., $z_{ij}(\cdot) = z_{ji}(\cdot)$.

We shall also assume:

(iv) $Z(\infty) = 0$, i.e., the degree of every numerator polynomial in $Z(\cdot)$ is less than the degree of the corresponding denominator.

Requirement (iv) is not really restrictive but it will simplify considerably the formulas which are to follow.

The matrix function $Z(\cdot)$ represents the input-output relations of the N-port. It is natural to examine the relation between this "external" description of the N-port and the "internal" description in terms of dynamical or state variables. That is, how does one associate a state with the N-port described in terms of its impedance matrix?

This question of representation has been settled recently by the writer (see particularly [1]). It turns out that every "transfer function" matrix $Z(\cdot)$ which has properties (i) and (iv), but not necessarily (ii) and (iii), one may associate a system of vector differential equations of the form

$$dx/dt = Fx + Gu(y), \tag{1.2}$$

$$y(t) = Hx(t). \tag{1.3}$$

Here x, the state, is an n-vector; $u(\cdot)$, the input (current) is an N-vector, and $y(\cdot)$, the output (voltage), is also an N-vector. F. G. H are constant linear transformations. We call (1.2 -3) a finite dimensional constant linear dynamical system [1].

---

*This research was supported in part by the US Air force under contracts AF33(657)-8559 (Aeronautical Systems Division) and AF 49(638)-1206 (Office of Scientific Research), and by the National Aeronautical and Space Administration under Contract NASw-718.

If equations (1.2-3) are known, the matrix Z(s) can be written down by inspection by taking the formal laplace transform of (1.2). The result is expressed by the formula

$$(1.4) \qquad Z(s) = H(sI - F)^{-1}G \quad (I = \text{unit matrix}).$$

Given Z(·), the determination of F, G, and H in (1.2-3) is much less trivial. Some set $\{F, G, H\}$ satisfying (1.4) always exists. Moreover, there is a smallest interger n_o such that relations (1.2-4) hold simultaneously. Generally speaking, this smallest dimension n_o is a complicated function of the matrix Z(·). If n in (1.2-3) is larger than n_o, then the dynamical system (1.2-3) is said to be *reducible*. The number n_o is identical with the so-called degree of Z(·) as defined by McMillan [2-4]. A numerical method for machine computing n_o was given in [1]. Alternately, McMillan's definition yield n_o via the so-called Smith canonical form of polynomial matrices [5].

From the mathematical point of view, equations (1.2-3) may be viewed as representing an *abstract* dynamical system defined with respect to an *abstract* vector space X. To give these equations concrete meaning, we must choose a specific coordinate system or *basis* in X and express the abstract vector x and the abstract linear transformations F, G, and H in numerical form. Once this has been done, x becomes an n-tuple of real numbers and F, G, H become n × n, n × N, and N × n arrays (matrices) of real numbers.

Any system (1.2-3) given in numerical form, as just described, is called a *realization* of Z(·) (see [1]). Mathematically, the term "realization" means that we pass from the abstract to the concrete (numerical) description. Physically, the term "realization" is motivated by the fact [5] that any numerically given system of equations (1.2-3) may be interpreted as the program for an analog computer which simulates the given N-port.

East realization corresponds to a specific choice of a coordinate system for the state vector. Our ultimate objective is to obtain that subclass of realizations which can be identified with a passive network, not merely with an analog computer program.

The next problem concerns the study of the relationships between various realizations of Z(·). This is indeed the main idea motivating the research discussed here. The problem is clearly of a *group theoretical* nature. We ask: What is the group of transformations which leave the properties of a given realization invariant?

Suppose we pick two bases for representing the abstract vector x. In the first, the vector is x described by the numerical n-tuple $\xi = (\xi_1, \ldots, \xi_n)$ and in the second it is represented by the n-tuple $\hat{\xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_n)$. It is well known [7, p. 82] that $\xi_1\hat{\xi}$ are related by a nonsingular linear transformation, so that

$$(1.5) \qquad \hat{\xi} = T\xi \text{ or } \hat{\xi}_i = \sum_{i=1}^{n} t_{ij}\xi_j, \quad i = 1, \ldots, n$$

where $T = [t_{ij}]$ is a constant, real, nonsingular matrix. The matrices T form the general linear group.

It is convenient to abuse notation and employ the symbols x, $\hat{x}$; F, $\hat{F}$, ... also for the n-tuples representing the vector x and matrices representing the linear transformation F with respect to certain specified bases. Now if the set $\{F, G, H\}$ specifies the dynamical system in the first basis, then the set $\{F, G, H\}$ which specifies the same system with respect to the second basis is related to the first set by the relations

$$\hat{F} = TFT^{-1}$$
$$\hat{G} = TG \tag{1.6}$$
$$\hat{H} = HT^{-1}$$

which are easily derived using (1.2-3) and (1.5). (See [1].). Different choices of bases correspond to different realizations of the same Z(·). Therefore one would certainly expect that Z(·) is invariant with respect to a change in basis. This is verified with the aid of (1.6):

$$\begin{aligned} Z(s) = H(sI - F)G &= HT^{-1}T(sI - F)^{-1}T^{-1}TG \\ &= \hat{H}(sI - TFT^{-1})^{-1}\hat{G} = \hat{H}(sI - \hat{F})^{-1}\hat{G}. \end{aligned}$$

Conversely, one may ask: In what way do any two realizations of Z(·) differ from one another? The answer is [1] that if they are irreducible ($n = n_o$), then they differ only by a choice of basis. The criterion for irreducibility is that the triple F, G, H be completely controllable and completely observable [1]. Thus if two completely controllable and completely observable triples $\{F, G, H\}$ and $\{\hat{F}, \hat{G}, \hat{H}\}$ yield the same transfer function matrix Z(·), then they are necessarily connected by the relations (1.6). Note that this is an abstract result; in practical cases it may be quite difficult to find the transformation T.

We now rephrase this important fact in such a way as to emphasize its group theoretical character:

THEOREM. Any two irreducible realizations (1.2-3) of a transfer function matrix Z(·) having properties (i) and (iv) are equivalent under the general linear group.

Now if Z(·) is the impedance matrix of a passive N-port, it will have certain other properties (namely (ii-iii) above) in addition to those needed to establish this theorem. One would therefore expect to find more restricted types of realizations which are invariant with respect to certain subgroups of the general linear group. The determination of these "network subgroups" is identical with the problem of studying all possible network realizations of a given Z(·), which is also called the problem of network equivalence. An important advantage of the group theoretic approach we wish to explore here is that it provides a unified way of studying the problem of synthesis by different classes of elements. For instance, the RLCT and RLCTT synthesis problems can be studied simultaneously. (See [8].)

It is important to bear in mind the conceptual distinction between the "impedance" transfer function and the "state variable" points of view in network theory.

Transfer and impedance functions are coordinate-free notions. They are most useful in studying properties of networks regardless of their internal structure. This is the deeper reason why existence criteria (such as positive realness) are stated more conveniently in terms of $Z(\cdot)$ than in terms of the triple $\{F, G, H\}$. This observation is not confined to network theory [9-10].

On the other hand, dynamical equations (1.2-3) always involve coordinates. These equations are most useful in the detailed study of the internal structure of a network. Such considerations have been generally absent from classical network theory, which may explain in part the difficulties encountered in resolving network equivalence problems.

The group-theoretical approach suggested here is completely analogous to the famous Erlanger Programm of Felix Klein. There have never been any systematic effort to apply Klein's ideas to network theory, as far as the writer is aware. It should be pointed out, however, that the work on network equivalence of Cauer [11, see particularly pages xviii, 49, and Chapter 10] was certainly a conscious step in the same direction. Much more can be done along these lines.

2. Restrictions Due to the Passivity and Reciprocity.

The fact that $Z(\cdot)$ represents a passive N-port imposes certain restrictions on the matrices F, G, and H. These restrictions are the counterpart of properties (ii-iii) of the impedance matrix.

First of all it is necessary to identify the components x_i of the state n-tuple x with physical variables in the network. We adopt the following convention, which is both standard and convenient [12].

Let us consider an N-port which contains n_L inductors and $n_C = n - n_L$ capacitors. Then we define

$$\begin{aligned} x_i &= \text{current through i-th inductor, when } i = 1, \ldots, n_L; \\ x_j &= \text{voltage across i-th capacitor, when } j = n_L + 1, \ldots, n. \end{aligned} \tag{2.1}$$

Assuming for a moment that none of the inductors are coupled with each other, and the same for the capacitors, it follows from linearity that the dynamical equations of the network may be written in the form

$$\begin{cases} L_i dx_i/dt = \sum_{k=1}^{n} a_{ik}x_k + \sum_{\ell=1}^{N} b_{i\ell}J_\ell(t) & i = 1, \ldots, n_L \\ C_j dx_j/dt = \sum_{k=1}^{n} a_{jk}x_k + \sum_{\ell=1}^{N} b_{j\ell}J_\ell(t) & j = n_L+1, \ldots, n \\ U_\ell(t) = \sum_{k=1}^{n} h_{\ell k}x_k(t), & \ell = 1, \ldots, N. \end{cases} \tag{2.2}$$

The $J_\ell(t)$ are the currents entering the ports and the $U_\ell(t)$ are the voltages across the ports.

To derive the numbers a_{ik} and $b_{i\ell}$ we may replace, for an instant, an inductor by a current source and a capacitor by a voltage source. We then obtain the following interpretation.

Let i_1, i_2 be integers belonging to $[1, n_L]$, j_1, j_2 integers belonging to $[n_L, n]$, and ℓ an integer belonging to $[1, N]$.

$a_{i_1 i_2}$ = voltage across i_1-th inductor when all capacitors are short circuited, all ports and all inductors save the i_2-th are open circuited, and the i_2-th inductor is replaced by a unit current source.

$a_{i_1 j_2}$ = voltage across the i_1-th inductor when all ports and all inductors are open circuited, all capacitors save the j_2-th are short circuited, and the j_2-th capacitor is replaced by a unit voltage source.

$b_{i_1 \ell}$ = voltage across the i_1-th inductor when all capacitors are short circuited, all inductors are open circuited, all ports save the ℓ-th are open circuited, and a unit current source is connected across with ℓ-th port.

The other quantities are defined analogously.

It is clear that the matrices A and B depend only on that part of the network which contains the resistors and ideal transformers. Partioning these matrices according to the numbering scheme introduced above, we can easily see what restrictions are imposed by passivity and/or reciprocity. If we write

(2.3)
$$A = \left[\begin{array}{c|c} A_1 & C_2 \\ \hline C_1 & A_2 \end{array}\right]$$

then A_1 has the dimension of resistance, A_2 has the dimension of conductance, while C_1 and C_2 are dimensionless.

Passivity implies that A_1 and A_2 are nonpositive (but not necessarily symmetric) matrices i.e., their quadratic froms are nonpositive. Moreover, the quadratic form

$$x' \begin{bmatrix} 0 & C_1 \\ C_2 & 0 \end{bmatrix} x$$

must be identically 0, i.e., $C_2 = -C_1'$.

Reciprocity implies that A_1 and A_2 are symmetric (but not necessarily nonpositive).

As far as B and H are concerned,

$$B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix}.$$

If $B_1 \neq 0$ then $Z(\infty) \neq 0$ and assumption (iv) of the previous section is violated. $H_1 = 0$ for the same reason. Passivity requires that H_2B_2 be nonnegative definite. Reciprocity requires $B_2 = H_2'$.

After removing all inductors and capacitors the N-port becomes an (N + n)-port. This (N + n)-port does not necessarily possess either an impedance or admittance matrix because there may be open circuits or short circuits at certain ports. In such cases, one adopts a suitable mixed impedance-admittance description of the last n ports. We do not wish to dwell upon the complications -- all trivial -- which result in such cases.

Suppose we write the energy stored in the N-port as

$$E = \frac{1}{2} x'Px = \frac{1}{2} \sum_{i=1}^{n_L} L_i x_i^2 + \frac{1}{2} \sum_{j=n_L+1}^{n} C_j x_j^2 .$$

Then equations (2.2) take the simple form

$$Pdx/dt = Ax + BJ(t), \quad U(t) = Hx(t). \tag{2.2'}$$

These equations are valid even if there is coupling between the inductors and capacitors. Thus in general P will have the form

$$P = \left[\begin{array}{c|c} L & 0 \\ \hline 0 & C \end{array}\right]. \tag{2.5}$$

The off-diagonal terms of P are 0 because in conventional networks there is no coupling between electric and magnetic fields. Since the stored energy must be a positive definite quadratic form, we assume that L and C are positive definite matrices. Reciprocity requires that they be symmetrical matrices. In the conventional RLCT cases L is symmetric, positive definite, while C is a matrix with positive entries on the diagonal and zeros elsewhere.

The case when L is merely nonnegative definite indicates the presence of ideal transformers. In this case the number of state variables is too large. We shall not discuss the resulting complications.

If the dynamical equations (2.2') of the N-port are reducible, i.e., may be replaced by a smaller set of equations having the same impedance matrix Z, then the network contains dynamical modes which are not specified by the impedance matrix Z but arise solely as a result of the synthesis procedure. For instance, the Darlington and the Bott-Duffin procedures introduce such extraneous modes. Although the presence of such additional modes may be necessary to carry out certain types of synthesis procedures, in this paper we shall be concerned only with the irreducible case. In other words, it will always be assumed that the N-port always contains a minimal number n_o of reactive elements.

3. Characterization of Passivity.

Now we shall state a relation between the impedance Z and the state variable description of an N-port. This relation was discovered in the course of studying the so-called Lur'e problem of constructing Lyapunov functions for dynamical systems which are linear save for a single nonlinear element [9].

The result to be stated below is more general than the Main Lemma in [9], in that we admit $N \times N$ rather than 1×1 impedance matrices and we drop the assumption that all eigenvalues of F have negative real parts. On the other hand, we will assume here, as a matter of convenience, that $Z(\infty) = 0$. This is an unessential restriction which was not needed in the Main Lemma of [9]. For a full treatment of the general problem, including proofs, see [13, 14].

CHARACTERIZATION OF PASSIVITY THEOREM. Let $Z(\cdot)$ be an $N \times N$ matrix of rational functions of the complex variable s, with $Z(\infty)= 0$. Let F, G, H be a triple such that (1.2-c) is an irreducible realization of $Z(\cdot)$. Let $\Psi(y)$ be a continuous p-vector function of the p-vector y such that $\Psi(0)= 0$ and $y'\Psi(y) \geqq$ for all y.

Then the following statements are equivalent:

(I) $Z(.)$ is nonnegative real, i.e., Re $s \geq 0$ implies $Z(s)+ Z'(\bar{s})=$ nonnegative definite hermetian matrix.

(II) There exists a symmetric, positive definite matrix P and a symmetric, nonnegative definite matrix Q such that

$$PF + F'P = - 2Q, \tag{3,1}$$

$$PG = H. \tag{3.2}$$

(A matrix F satisfying (3.1-2) cannot have an eigenvalue with positive real part; in its Jordan form all imaginary eigenvalues are contained in 1×1 blocks; the null space of the matrix Q is necessarily contained in the eigenspace of F spanned by the eigenvectors corresponding to imaginary eigenvalues.)

(III) $V = x'Px$ is a Lyapunov function for the autonomous dynamical system $dx/dt = Fx - G\Psi(Hx)$ such that $\dot{V}(x) \leqq 0$.

Let us give an indication of the proof. (II) implies (I) by direct substitution. Given (I), the nonnegative real character of $Z(i\omega)$ allows it to be factored as $Z(i\omega) + Z'(-i\omega) = W'(i\omega)W(i\omega)$. Every such factorization gives rise to a Q. The irreducibility of the realization is then utilized to verify that (3.1-2) hold. (II) implies (III) because

$$\dot{V}(x) = 2x'PFx - 2x'G'P\Psi(Hx)$$

$$= - 2\,[x'Qx + y'\Psi(y)\,] \leqq 0$$

by (3.1-2). Finally, (III) implies (I) by direct computation.

Remarks. (1) In a 1958 paper, Desoer [15] pointed out that at that time the stability of passive nonreciprocal networks has not yet been proved and gave an argument showing that all such networks are, in fact, (Lyapunov) stable. This result is now confirmed in the strongest possible way. Since the nonnegative realness of Z(·) (as stated above, i.e., not requiring Z(·) = Z'(·)) is a necessary condition for realizing a passive RLCTT N-port, it is clear from the parenthetical remark in (II) that F cannot have eigenvalues with positive real parts, nor can it have solutions of the type $t^k \cos(\omega t + \alpha)$, $k > 0$.

(2) For networks, V = x'Px can always be identified with the stored energy in the inductors and capacitors. Hence (III) shows that in unterminated passive networks ($\Psi = 0$) the energy is a nonincreasing function of time. For arbitrary passive resistive terminations, which may be either linear or nonlinear, the same conclusion holds. In the nonlinear case this result represents a considerable improvement over what was know before. (To my knowledge, the previous best general result here is that of Duffin [16].)

The characterization theorem expresses facts which are usually taken for granted from an intuitive physical point of view. As a result, the average engineer may be tempted to jump to the conclusion that nothing new has been done. But to do this would be a gross misunderstanding of the processes of scientific research. The theorem is a more precise and more general characterization of passivity than was heretofore available. The precision and the generality of this result in turn leads to deep insight into problems of network synthesis [8, 14], which can now be studied with a simplicity and explicitness that was impossible previously.

This theorem is a very convincing additional piece of evidence that passivity and network synthesis are intimately related. In [8] I shall present some preliminary results which show that the structural properties of the class of all networks which realize a given impedance function Z(·) are in 1-1 correspondence with certain algebraic invariants of Z(·).

4. Applications.

We shall illustrate the usefulness of the characterization theorem and the group theoretic ideas related to it by two classical examples: (i) LC 1-port synthesis according to Foster, and (ii) synthesis of lossless 2-ports.

EXAMPLE 1. We wish to prove that

A necessary and sufficient condition in order that the scalar impedance function z(·) be realizable as the 1-port shown in Fig. 1 is that

(i) $z(\cdot)$ is nonnegative real and

(ii) all poles of $z(\cdot)$ are imaginary.

Necessity: If the state variables for the network are assigned as shown in Fig. 1, it follows by inspection that the A, B, H, and P matrices are

$$A = \left[\begin{array}{c:cc:c:cc} 0 & & & & & \\ \hdashline & 0 & 1 & & & \\ & 1 & 0 & & & \\ \hdashline & & & \ddots & & \\ \hdashline & & & & 0 & 1 \\ & & & & -1 & 0 \end{array}\right] \tag{4.1}$$

There are r 2 × 2 blocks on the diagonal

$$B = \left[\begin{array}{c} 1 \\ \hdashline 0 \\ 1 \\ \hdashline \vdots \\ \hdashline 0 \\ 1 \end{array}\right] \tag{4.2}$$

$$H = [1 \;\; 0 \;\; 1 \;\; . \; . \; . \;\; 0 \;\; 1] \tag{4.3}$$

$$P = \left[\begin{array}{c:cc:c:cc} C_o & & & & & \\ \hdashline & L_1 & 0 & & & \\ & 0 & C_1 & & & \\ \hdashline & & & \ddots & & \\ \hdashline & & & & L_r & 0 \\ & & & & 0 & C_r \end{array}\right] \tag{4.4}$$

It is clear that the quintuple

$$F = P^{-1}A, \; G = P^{-1}B, \; H = H, \; P = P, \; Q = 0$$

satisfies (3.1-2). Hence the impedance function of the 1-port, which is given by

$$z(s) = \frac{1}{C_o s} + \sum_{k=1}^{r} \frac{s}{C_k(s^2 + \omega_k^2)}, \qquad \omega_k^2 = 1/L_k C_k \tag{4.5}$$

is a nonnegative real function.

Sufficiency: Suppose that $z(\cdot)$ is nonnegative real. Then relations (3.1-2) are satisfied. If, in addition, all poles of $z(\cdot)$ are imaginary, then Q in (3.1) is necessarily zero.

We shall now carry out a series of changes of basis which will eventually exhibit the matrices $A = PF$, $B = PG$, $H = H$, and $P = P$ in the form (4.1-4), after which the existence of the network shown in Fig. 1 is obvious by inspection.

Step 1. We pick a basis that $P_{(1)} = I$. Then by (3.1) $F_{(1)} = -F'_{(1)}$. This is always possible because under a change of basis (1.5) we have

$$\text{(1.6')} \qquad P = T'\hat{P}T$$

Since P is positive definite, it can always be written as $P = T'IT$, T nonsingular, so that $\hat{P} = P_{(1)} = I$.

Step 2. We pick a second basis which leaves P invariant ($P_{(2)} = I$) but transforms $F_{(1)}$ into the canonical form

$$F_{(2)} = \left[\begin{array}{c|cc|c|cc} 0 & & & & & \\ \hline & 0 & \omega_1 & & & \\ & -\omega_1 & 0 & & & \\ \hline & & & \ddots & & \\ \hline & & & & 0 & \omega_r \\ & & & & -\omega_r & 0 \end{array}\right] .$$

This is always possible because [17] every skew symmetric matrix can be transformed into the canonical form (4.6) by means of an orthogonal transformation. Note that the element 0 in the upper left-hand corner of $F_{(2)}$ corresponds to the eigenvalue 0 (if $F_{(1)}$ happens to have that eigenvalue), while the 2 × 2 blocks along the diagonal correspond to the eigenvalues $\pm i\omega_k$, $k = 1, \ldots, r$. (Since $z(\cdot)$ has only imaginary poles, F has only imaginary eigenvalues; by nonnegative realness all eigenvalues must be simple. Thus $F_{(2)}$ is as it should be.)

In the basis (2), the column vectors B, H' have the form

$$G_{(2)} = B_{(2)} = H'_{(2)} = \left[\begin{array}{c} \beta_o \\ \hline \alpha_1 \\ \beta_1 \\ \hline \cdot \\ \cdot \\ \cdot \\ \hline \alpha_r \\ \beta_r \end{array}\right]$$

Since the triple $\{F, G, H'\}$ must be irreducible and therefore completely controllable and completely observable if (II) of the characterization theorem is to hold, it is easily verified that $\alpha_o > 0$ and $\alpha_k^2 + \beta_k^2 > 0$, $k = 1, \ldots, r$.

Step 3. Any proper (det = 1) orthogonal transformation (rotation) applied to the 2×2 matrix

$$\begin{bmatrix} 0 & \omega_k \\ -\omega_k & 0 \end{bmatrix}$$

leaves it invariant. We apply r such rotations corresponding to the r blocks in F, in such a way that the vectors

$$\begin{bmatrix} \alpha_k \\ \beta_k \end{bmatrix}, \qquad k = 1, \ldots, r$$

in B and H' are rotated into the vectors

$$\begin{bmatrix} 0 \\ \gamma_k \end{bmatrix}, \qquad k = 1, \ldots, r.$$

By the comment made at the end of the preceding paragraph we know that $\gamma_k = \alpha_k^2 + \beta_k^2 \neq 0$. Thus $F_{(3)} = F_{(2)}$, $P_{(3)} = P_{(2)} = I$ and

$$G_{(3)} = B_{(3)} = \left[\begin{array}{c} \beta_o \\ \hline 0 \\ \gamma_1 \\ \hline \vdots \\ \hline 0 \\ \gamma_r \end{array}\right].$$

Step 4. Now we obtain change to a basis where A, B and H have only 0, 1, -1 as elements. Fortunately this can be accomplished with the aid of a diagonal matrix. If $x_{(3)} = \Lambda x_{(4)}$, where Λ is diagonal, then

$$P_{(4)} = \Lambda P_{(3)} \Lambda = \Lambda^2,$$

$$B_{(4)} = P_{(4)} G_{(4)} = \Lambda P_{(3)} \Lambda \Lambda^{-1} G_{(3)} = \Lambda B_{(3)}$$

and

$$A_{(4)} = P_{(4)} F_{(4)} = \Lambda P_{(3)} \Lambda \Lambda^{-1} F_{(3)} \Lambda = \Lambda A_{(3)} \Lambda .$$

We choose

$$\Lambda = \left[\begin{array}{c|cc|c|cc}
1/\beta_o & & & & & \\
\hline
 & \gamma_1/\omega_1 & 0 & & & \\
 & 0 & 1/\gamma_1 & & & \\
\hline
 & & & \ddots & & \\
\hline
 & & & & \gamma_r/\omega_r & 0 \\
 & & & & 0 & 1/\gamma_r
\end{array}\right].$$

It is easily checked that $A_{(4)}$, $B_{(4)}$, $H_{(4)}$ assume the form (4.1 - 4.3). If we let

$$C_o = 1/\beta_o,$$
$$C_k = 1/\gamma_k, \qquad k = 1, \ldots, r;$$
$$L_k = \gamma_k/\omega_k, \qquad k = 1, \ldots, r;$$

then the same can be said also about $P_{(4)}$.

The proof is completed.

It may appear to the reader well-versed in classical synthesis theory that the usual Foster solution, obtained by a partial fraction expansion of (4.5), is a much simpler road to obtaining the network. Actually, the total number of arguments is not really smaller than in the matrix case. Our present method, however, has the very important additional feature that it gives insight into the group structure of passive 1-ports. Only a few more arguments are needed to completely describe in this way all passive 1-ports which contain a minimal number of reactive elements. The partial fraction manipulations used in the Foster theory can be interpreted as calculations based on the group representation provided by the laplace transform.

EXAMPLE 2. Let us consider the impedance matrix

$$Z(\cdot) = \begin{bmatrix} \dfrac{k_{11}s}{s^2+1} & \dfrac{k_{12}s}{s^2+1} \\[2ex] \dfrac{k_{12}s}{s^2+1} & \dfrac{k_{22}s}{s^2+1} \end{bmatrix}.$$

We wish to realize this matrix with a (lossless) reciprocal 2-port containing a minimum number of reactive elements.

Reciprocity requires that $k_{21} = k_{12}$. Passivity requires (as may be easily checked using the characterization theorem)

$$\Delta = k_{11}k_{22} - k_{12}^2 \geq 0,$$

which is popularly known as the "residue condition". In fact, the matrix of residues of $Z(\cdot)$ at either pole $s = \pm i$ is given by

$$K = \begin{bmatrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{bmatrix}$$

and $\Delta = \det K$.

Two possibilities may arise: Either $\Delta > 0$ or $\Delta = 0$.

The minimum number of reactive elements to be used is equal to the minimum number n_o of state variables which are necessary and sufficient to realize $Z(\cdot)$ as a dynamical system (1.2-3). The number n_o can be computed with the aid of a theorem of E. G. Gilbert [Theorem 11], which states (in the special case when $Z(\cdot)$ has distinct poles)

$$n_o = \Sigma \text{ ranks of residue matrices of } Z(\cdot) .$$

Hence $n_o = 4$ when $\Delta > 0$ and $n_o = 2$ when $\Delta = 0$. It should be noted that n_o is also equal to the McMillan degree of $Z(\cdot)$ [4].

Let us consider the case $\Delta > 0$. In [1, Sect. 8] two methods were given for constructing realizations of a given transfer function matrix. Using Method (B), it is found that the following matrices provide a realization of $Z(\cdot)$:

$$F_{(1)} = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}, \quad G_{(1)} = \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad H_{(1)} = [0 \quad K] .$$

If we let

$$P_{(1)} = \begin{bmatrix} K & 0 \\ 0 & K \end{bmatrix}, \quad Q_{(1)} = 0,$$

then the quintuple $\{F_{(1)}, G_{(1)}, H_{(1)}, P_{(1)}, Q_{(1)}\}$ satisfies (3.1-2). However, these matrices do not correspond to a network realization of $Z(\cdot)$. In order to obtain such a realization, we introduce a new basis defined by $x_{(1)} = Ux_{(2)}$, where

$$U = \begin{bmatrix} I & 0 \\ 0 & K^{-1} \end{bmatrix} .$$

Then

$$A_{(2)} = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}, \quad B_{(2)} = \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad P_{(2)} = \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} .$$

Referring to the discussion of Sect. 1, it is easily seen that these matrices represent two resonant LC circuits in which there is coupling (represented by the matrices K and K^{-1} respectively) between both the inductors and capacitors.

Since coupled capacitors can always be realized by ordinary capacitors and ideal transformers, we have proved:

When $\Delta > 0$, $Z(\cdot)$ can be realized as an LCT network with a minimal number of reactive elements.

By further easy manipulations of the matrices involved, it can be shown -- as is well known [19, p. 447-453] -- that a single ideal transformer will always suffice.

The question then arises whether or not the 2-port can be realized using only coupled inductors (but not coupled capacitors) and without ideal transformers; further, whether or not the 2-port can be realized without any inductive or capacitive coupling. The answer to both questions is in general no. We shall not give the proof (straightforward) but merely state the result:

When $\Delta > 0$, $Z(\cdot)$ can be realized as an LC network with a minimal number of reactive elements, without ideal transformers, and without any coupling between inductors or capacitors if and only if either (i) $k_{12} = 0$, or (ii) $k_{12} = k_{11}$, or (iii) $k_{12} = k_{22}$. These conditions cannot be weakened if either coupled inductors or coupled capacitors (but not both) are admitted.

In the first case the 2-port consists of two unconnected 1-ports. In the second and third case the 2-port can be realized as an L-section.

I don't know how well known this easy result is. It is quite useful theoretically because it shows that transformerless synthesis implies very strong constraints on $Z(\cdot)$ in addition to nonnegative realness. This fact has a direct bearing on the transformerless synthesis of RLC 1-ports because such problems can always be reduced to the synthesis of lossless N-ports. (See [8].)

The case $\Delta = 0$ may be treated similarly.

5. Conclusions.

(1) Methods based on the impedance concept are coordinate-free. They do not display directly the structural properties of the realization.

(2) State-variable methods on the other hand, are closely related to the structural properties of networks.

(3) Every transfer function matrix admits realizations. Some of these realizations may correspond to networks, while others may require active elements (analog computers).

(4) Nonnegative real impedance matrices always admit passive realizations, i.e. realizations impose restrictions on F, G, H, P, Q in addition to those implied by passivity. In all cases that these restrictions are known they are expressible in terms of the impedance matrix.

To classify all types of realizations of a given abstract dynamical system and to express the realizability conditions as coordinate-free properties (*such as passivity*, *reciprocity*, *etc.*) *of abstract dynamical systems*.

This problem suggests a partnership between mathematics and network theory which will be intellectually exciting and practically profitable.

Dynamical systems are the building blocks of modern technology. The resolution of the problem posed will tell us what technology is capable of doing at present. It will also suggest developing new components to realize that which is known to be possible for mathematical reasons.

6. *References*.

[1] *R. E. Kalman*, "Mathematical description of linear dynamical systems", SIAM J. Control, *1* (1963) 152-192.

[2] *Brockway McMillan*, "Introduction to formal realizability theory", Bell System Technical Journal, *31* (1952) 217-279, 541-600.

[3] *M. Bayard*, THÉORIE DES RÉSEAUX DE KIRCHHOFF (book), Paris, 1954.

[4] *R. E. Kalman*, "Irreducible realizations and the degree of a matrix of rational functions", to appear.

[5] *F. R. Gantmakher*, THEORY OF MATRICES (book, 2 volumes), Chelsea, 1959.

[6] *R. E. Kalman*, "Analysis and design principles of second and higher order saturating servomechanisms", Trans.AIEE, *74*, II (1955) 293-310.

[7] *P. R. Halmos*, FINITE-DIMENSIONAL VECTOR SPACES (book), Van Nostrand,1958.

[8] *R. E. Kalman*, "A new look at the RLCTT synthesis problem", Proc. First Allerton Conf. on Circuit and System Theory, 1963.

[9] *R. E. Kalman*, "Lyapunov functions for the problem of Lur'e in automatic control", Proc. Nat.Acad.Sci. (USA), *49*, (1963) 201-205.

[10] *R. E. Kalman*, "When is a linear control system optimal?", J. Basic Engr. 1964 (to appear).

[11] *Wilhelm Cauer*, SYNTHESIS OF LINEAR COMMUNICATION NETWORKS (book), McGraw-Hill, 1958.

[12] *R. E. Kalman*, "Network analysis by the state-variable method", to appear.

[13] *R. E. Kalman*, "Resolution of the problem of Lur'e", to appear.

[14] *R. E. Kalman*, "Synthesis of RLCT N-ports", to appear.

[15] *C. A. Desoer*, "On the characteristic frequencies of lossless nonreciprocal networks",Trans.IRE PGCT, *5*, (1958) 374-375.

[16] *R. J. Duffin*, "Nonlinear networks I", Bull.Am.Math. Soc. *52* (1946) 833-838.

[17] See Chapter 11, Volume II of Reference 5.

[18] *R. E. Kalman*, "On the structure of lossless N-ports", to appear.

[19] *Louis Weinberg*, NETWORK ANALYSIS AND SYNTHESIS (book), McGraw-Hill,1962.

SOME RESULTS IN LINEAR MULTIPLE LOOP FEEDBACK SYSTEMS

E. S. KUH
Electronics Research Laboratory
University of California
Berkeley, California

I. INTRODUCTION

It is well-known that the concept of return difference plays an important role in the theory of linear feedback systems. First, it is pertinent in the determination of stability. The fact that return difference (or loop transmission) can be easily measured in many practical feedback systems reduces the problem of stability study to a Nyquist plot. Second, sensitivity, which is a measure of the change of over-all behavior with respect to change of a particular element in the system, can be expressed in terms of return difference. In many situations, sensitivity is approximately equal to the reciprocal of the return difference. Third, return difference can be used in determining the driving-point and transfer functions of a system.

Recent work by Goldwyn[1] and Sandberg[2] represents important generalization of Bode's[3] theorems on return difference to multiple loop feedback systems. Sandberg considers the problem where a matrix control source X (instead of a scalar control element x) is of main concern in the system. He proceeds to introduce a loop difference matrix for X and derives many interesting theorems in terms of properties of the determinant of the loop difference matrix. In particular, he generalizes Blackman's[4] impedance formula and obtains the following theorem: If $w(0) \neq 0$

$$w(X) = w(0) \frac{|\hat{F}(X)|}{|F(X)|} \tag{1}$$

where $w(X)$ is a scalar system function, $w(0)$ is the function w when X is set equal zero, and $F(X)$ and $\hat{F}(X)$ are the loop difference and the null loop difference matrices for the control source X. The notation $|F(X)|$ represents the determinant of the matrix $F(X)$.

In this paper we deal with linear feedback systems which contain a multiplicity of inputs, outputs, and feedback loops. This is

The research reported herein is made possible through support received from the Departments of Army, Navy, and Air Force under grant AF-AFOSR-139-63.

obviously of great importance in view of present-day interests in multivariable control systems and linear n-port active circuits. We generalize directly Bode's return difference concept for single control source and single bilateral element to multivariable problems. We derive theorems which are extensions of Sandberg's theorems on loop difference matrices. Interpretation of return difference for an n-port admittance matrix X will be given along with that of a control source matrix X. Finally the problem of sensitivity will be discussed. Relations between sensitivity and the return difference matrix will be derived.

II. SYSTEM DESCRIPTION AND THE TRANSMITTANCE MATRIX, W(X)

In a feedback system there are three variable which are of major interest, namely: the input, the output, and a particular branch of interest in the system. Let the input variable be represented by an n-dimensional vector y and the output variable be represented by an m-dimensional vector z, and the branch of interest be a rectangular matrix X of dimension q x p. X can either be a transfer function matrix or a driving-point function matrix described by the equation

$$\psi = X\phi . \tag{2}$$

If X is a transfer function matrix, ϕ represents a p-dimensional controlling variable vector and ψ represents a q-dimensional controlled variable vector. A typical example of a 2 x 2 matrix X is shown by the vacuum tube circuit of Fig. 1. The elements in the matrix X are then the transconductances of the four pentodes. The controlling variables are the voltages ϕ_1 and ϕ_2 at the input, while the controlled variables are the currents ψ_1 and ψ_2 at the output.

If X represents a driving-point function matrix, ϕ and ψ are of same dimension, p = q, and they represent voltage and current vectors of a p-port network.

Eq. (2) describes the relation for the branch X which represents part of the over-all system. The branch X is important in terms of its effects to the whole system and is imbedded in the rest designated by H as shown in the block diagram of Fig. 2. The properties of the over-all system in terms of X, such as sensitivity and stability are of main concern to system designers. Consider now the block diagram H by itself. The vectors y and ψ are then inputs to H, while vectors z and ϕ are independent outputs of H. Since the system is linear, we can express ϕ and z in terms of y and ψ as follows:

$$\phi = Ay + C\psi \tag{3}$$

$$z = Dy + B\psi \tag{4}$$

where A, B, C, and D are transmittance matrices of order p x n, m x q, p x q, and m x n, respectively. The over-all transmittance matrix which relates the input and output are defined by

$$z = W(X)\, y \tag{5}$$

where W(X) is of dimension m x n. Combining (2), (3), (4), and (5), we obtain

$$W(X) = D + BX\,(1_p - CX)^{-1}A \tag{6}$$

or

$$W(X) = D + B(1_q - XC)^{-1}XA \tag{7}$$

and

$$W(0) = D. \tag{8}$$

Eq. (6) and (7) express the over-all transmittance of the feedback system in terms of simpler transmittances of the constituents in the system with known properties.

The above can also be represented by a vector signal flow graph as shown in Fig. 3. Using the rules of flow graph reduction, we can easily obtain (6) and (7).

III. THE RETURN DIFFERENCE AND THE NULL DIFFERENCE MATRICES FOR X

The classical return difference of Bode is defined in terms of the network determinant and cofactor. It has the physical interpretation as one minus the loop transmission. In a single control source $I = xV$ as shown in Fig. 4, the loop transmission, $\mu\beta$, is defined as the return voltage V_3' at the input of the control source when the dependent source is replaced by an independent source of x amperes and when the excitation I_s is removed. It can also be interpreted in another fashion as shown in Fig. 5. Let the control source be broken off at the input and let a voltage source of g volt be applied to the right of the broken point. The return voltage V_3' to the left of the broken point is then g times the loop transmission. For a bilateral 1-port admittance x at port 3 in Fig. 6, the loop transmission can be interpreted in terms of the voltage at port 3 when the admittance x is replaced by an independent current source of x g amperes. The loop transmission is simply the voltage V_3' divided by g and the return difference F(x) has therefore the physical significance of the ratio of the admittance x + Y and the admittance Y, where Y is the admittance which x faces.

In the case of matrix control source X, the return difference matrix can therefore be defined in terms of a matrix loop transmission. If the input of the control source is broken up and a vector input g of p dimension is applied to the right of the broken point, the return signal vector h to the left of the broken point due to g is then defined as the product of the loop transmission and g. Thus the return difference matrix F(X) is defined according to the following equation

$$F(X)\, g = g - h . \tag{9}$$

Similarly, for an admittance matrix X of a p-port, the loop transmission can be defined in the same fashion, if the admittance matrix X is replaced by a vector current source of X g. The voltage appearing at the p-port is h and F(X) g = g - h. Since the voltage can be related to the admittance matrix Y which X faces, we have

$$h = - Y^{-1} X g. \tag{10}$$

Thus,

$$F(X) = 1_p + Y^{-1} X = Y^{-1} (Y + X) . \tag{11}$$

This is therefore a direct generalization of the return difference in the classical sense for a bilaterial admittance x.

The vector signal flow graph representation of the above can be interpreted similarly. With reference to Fig. 7, the branch X is broken up as shown. Let a vector input g be applied to the right of the broken point, the returned vector h is then CXg, therefore from (9)

$$F(X) = 1_p - CX . \tag{12}$$

We can similary define the null return difference matrix as in the classical situation of the null return difference function. With reference to Fig. 7, the excitation vector y is adjusted such that the output z is zero. In terms of the flow graph representation the necessary relation between g and y is given by

$$Dy = - BXg$$

or (13)

$$y = - D^{-1}Bxg.$$

We observe that (13) is possible only when D is nonsingular. The output z must therefore be of same dimension as the input y. Physically this is rather obvious since a unique adjustment is only possible if the number of entries for z and y are the same, i. e., $m = n$. With inputs y and g, h is found to be $Ay + CXg$. Using (13), we find the null return difference matrix as

$$\hat{F}(X) = 1_p - CX + AD^{-1}BX = 1_p - \hat{C}X \tag{14}$$

where

$$\hat{C} = C - AD^{-1}B. \tag{15}$$

The following generalization of Blackman's impedance formula can be stated.

Theorem: If $W(0) = D$ is nonsingular, then

$$|W(X)| = |W(0)| \frac{|\hat{F}(X)|}{|F(X)|} \tag{16}$$

Proof: From (6)

$$W(X) = W(0)\left[1_n + D^{-1}BX(1_p - CX)^{-1}A\right]$$

$$|W(X)| = |W(0)| \; |1_p + AD^{-1}BX(1_p - CX)^{-1}|$$

$$= |W(0)| \; |1_p - CX + AD^{-1}BX| \; |1_p - CX|^{-1}$$

$$= |W(0)| \; \frac{|\hat{F}(X)|}{|F(X)|}.$$

Corollary I: If $p = n$ and $A = 1_n$, then

$$W(X) = W(0)\,\hat{F}(X)\,F^{-1}(X). \tag{17}$$

Corollary II: If $q = n$ and $B = 1_n$, then

$$W(X) = W(0)\, F^{*-1}(X)\, \hat{F}^{*}(X) \tag{18}$$

where

$$F^{*}(X) = 1_q - XC$$

and

$$\hat{F}^{*}(X) = 1_q - X\hat{C}\,.$$

The above corollaries are easily obtained from the proof of the theorem in (16). The flow graph representations for the above assumptions are illustrated in Fig. 8.

As an application of the theorem in (16), let $W(X)$ be the impedance matrix of an n-port which imbeds a p-port network specified by its admittance matrix X as shown in Fig. 9. From (11)

$$F(X) = 1 + Z_{oc}X \tag{19}$$

$$\hat{F}(X) = 1 + Z_{sc}X \tag{20}$$

where Z_{oc} and Z_{sc} are respectively the open and short circuit impedance matrices which X faces. Eq. (16) states that

$$|W(X)| = |W(0)| \frac{|1 + Z_{sc}X|}{|1 + Z_{oc}X|}\,. \tag{21}$$

This is the direct extension of the familiar formula of the input impedance of a terminated 2-port.

IV. THE RETURN DIFFERENCE AND THE NULL RETURN DIFFERENCE MATRICES FOR A GENERAL REFERENCE

Bode's classical return difference for a general reference k can again be extended to the matrix case. In defining the loop transmission for a scalar control source x with respect to reference k, one replaces part of the control source $x' = x - k$ by an independent current source of x' ampere. For the matrix case, we refer to Fig. 10, where X is split into K and X' and only X' is broken up in the calculation of loop transmission. The signal flow graph of Fig. 10 can be redrawn to emphasize the effect of X' as shown in Fig. 11. The transmittances of branches in Fig. 11 are related to that of the original graph by the following equations:

$$\psi' = X'\phi' = (X - K)\phi' \tag{22}$$

$$C' = (1_p - CK)^{-1}C \tag{23}$$

$$A' = (1_p - CK)^{-1}A \tag{24}$$

$$B' = B(1_q - KC)^{-1} \tag{25}$$

$$D' = D + BK(1_p - CK)^{-1}A. \tag{26}$$

Similar to the return difference matrix, we break the branch X' and introduce an input vector g to calculate the return vector h. Thus, the return difference matrix for the branch X with respect to a reference K is defined according to

$$F_K(X)\, g = g - h$$

for the signal flow graph in Fig. 11. Thus

$$F_K(X) = 1_p - C'X'. \tag{27}$$

The null return difference matrix for the branch X with respect to the reference K is

$$\hat{F}_K(X) = 1_p - \hat{C}'X' \tag{28}$$

where

$$\hat{C}' = C' - AD'^{-1}B'.$$

The following theorems can now be stated:

Theorem:

$$F_K(X) = F^{-1}(K)\, F(X) \tag{29}$$

$$\hat{F}_K(X) = \hat{F}^{-1}(K)\, \hat{F}(X). \tag{30}$$

Proof: The proof of (29) is immediate if (22) and (23) are substituted in (27). Thus,

$$F_K(X) = 1_p - C'X' = 1_p - (1_p - CK)^{-1}C(X - K)$$

$$= (1_p - CK)^{-1} \left[1_p - CK - CX + CK \right]$$

$$= (1_p - CK)^{-1} (1_p - CX)$$

$$= F^{-1}(K)\, F(X).$$

The proof of (30) is straightforward but slightly involved and is omitted here.

Similar to the theorem given in (16) we can state the following theorem:

Theorem: If $W(K)$ is nonsingular, then

$$|W(X)| = |W(K)| \; \frac{|\hat{F}_K(X)|}{|F_K(X)|}. \tag{31}$$

Proof: From (16) we have

$$|W(X)| = |W(0)| \; \frac{|\hat{F}(X)|}{|F(X)|}$$

and

$$|W(K)| = |W(0)| \; \frac{|\hat{F}(K)|}{|F(K)|}.$$

Combining the above,

$$|W(X)| = |W(K)| \; \frac{|\hat{F}(X)|}{|\hat{F}(K)|} \; \frac{|F(K)|}{|F(X)|}.$$

Using (29) and (30), we obtain immediately (31).

The interpretation of the return difference matrix and the null return difference matrix for a general reference of the circuit shown in Fig. 9 are simply

$$F_K(X) = (1 + Z_{oc}K)^{-1} (1 + Z_{oc}X) \tag{32}$$

$$\hat{F}_K(X) = (1 + Z_{sc}K)^{-1} (1 + Z_{sc}X). \tag{33}$$

From (31),

$$|Z| = |Z(K)| \frac{|1 + Z_{sc}X| \quad |1 + Z_{oc}K|}{|1 + Z_{oc}X| \quad |1 + Z_{sc}K|} . \qquad (34)$$

Eqs. (29) and (30) can also be extended to cover more general situations. Let $K = X_1$ where X_1 represents the reference condition for X and is obtained from X by setting certain elements in X to a particular set of value, then from (29) we have

$$F(X) = F(X_1)\, F_{X_1}(X_2) .$$

Consider next X_1 and a reference matrix X_2, where X_2 is obtained from X_1 by setting certain elements in X_1 to a particular set of value, we have

$$F(X_1) = F(X_2)\, F_{X_2}(X_1).$$

Continuing this process, we can express $F(X)$ in a general form. The following expansion theorem for the return difference matrix is obtained:

Theorem: The return difference matrix can be expressed in the following finite product form,

$$F(X) = F(X_n) \prod_{i=n}^{1} F_{X_i}(X_{i-1}) \qquad (35)$$

where X_i is obtained from X_{i-1} for $i = 1$ to n.

V. SENSITIVITY

Sensitivity for a scalar system function $w(x)$ with respect to a scalar element x is defined as

$$S = \frac{x}{w} \frac{dw}{dx} = x \frac{d \log w}{dx} . \qquad (36)$$

In our problem since both the transmittance W and the branch X are matrices, we have various ways of measuring sensitivity.

In the first method we adopt the conventional definition and define sensitivity for a scalar element w in W in terms of an element x_{ij} in X. From the familiar result between sensitivity and the scalar return difference, we have

$$S^{w}_{x_{ij}} = \frac{1}{F} - \frac{1}{\hat{F}}$$

$$= \frac{1}{F}\left[1 - \frac{w(0)}{w(x_{ij})}\right] \tag{37}$$

where F and $\hat{F}$ are the scalar return difference and null return difference for the element x_{ij}. $w(x_{ij})$ and $w(0)$ are respectively the scalar system functions when x_{ij} has its normal value and x_{ij} vanishes. To obtain a relation between sensitivity and the return difference matrix, we proceed as follows: Let K be the reference matrix which is obtained from X by setting x_{ij} zero. Thus $X' = X - K$ has the form that the only nonzero element in the matrix is x_{ij} at the i-th row and j-th column. It is only necessary to demonstrate that the return difference matrix determinant $|F_K(X)|$ for the branch X with respect to the element K is identical to the classical return difference F for the element x_{ij}. Referring to our definition of the return difference matrix $F_K(X)$, we break the branch X'; and assign a particular input vector g of p-dimension of the form

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ g_j \\ 0 \\ 0 \end{bmatrix} \quad \text{j-th row}$$

as shown in Fig. 12.

The vector h due to g is therefore

$$h = C'X'g = \left[\begin{array}{ccccc} 0 & 0 & - & c'_{1i}\,x_{ij} & 0 \\ 0 & 0 & - & c'_{2i}\,x_{ij} & 0 \\ 0 & 0 & - & c'_{3i}\,x_{ij} & 0 \\ \hdashline 0 & 0 & - & c'_{ji}\,x_{ij} & 0 \\ \hdashline 0 & 0 & - & c'_{pi}\,x_{ij} & 0 \end{array}\right] \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ \hdashline g_j \\ \hdashline 0 \end{array}\right]$$

The return difference matrix $F_K(X)$ is

$$F_K(X) = 1_p - C'X'$$

and

$$|F_K(X)| = 1 - c'_{ji}\, x_{ij}$$

which is the classical return difference for the element x_{ij}. This is evident since g_j represents the input for the element x_{ij} and $x_{ij}\, g_j$ is then the independent source and c'_{ji} is the returned transmission for the calculation of the loop transmission h_j/g_j. From (37) we have

$$S^{w}_{x_{ij}} = \frac{1}{|F_K(X)|}\left[1 - \frac{w(0)}{w(x_{ij})}\right] \tag{38}$$

$$= \frac{|F(K)|}{|F(X)|}\left[1 - \frac{w(0)}{w(x_{ij})}\right]$$

$$= \frac{|F(X)|_{x_{ij}=0}}{|F(X)|}\left[1 - \frac{w(0)}{w(x_{ij})}\right] \tag{39}$$

A sensitivity matrix for W can therefore be defined as

$$S^{W}_{x_{ij}} = \frac{|F(X)|_{x_{ij}=0}}{|F(X)|}\, W' \tag{40}$$

where W' is a matrix of $m \times n$ and contains

$$\frac{w_{mn}(x_{ij}) - w_{mn}(0)}{w_{mn}(x_{ij})}$$

as a typical element at the m-th row and n-th column.

Next if a scalar measure of sensitivity for the whole system $W(X)$ with respect to x_{ij} is desired, we can sometimes adopt the following definition by using the norm of a matrix defined as

$$\|W(X)\| = \sum_{m,\,n} |w_{mn}| \tag{41}$$

and

$$\left\|\frac{dW}{dx_{ij}}\right\| = \sum_{m.\,n} \left|\frac{dw_{mn}}{dx_{ij}}\right| . \tag{42}$$

A scalar sensitivity is then defined as

$$S_{x_{ij}} = \frac{|x_{ij}|}{\|W\|} \left\| \frac{dW}{dx_{ij}} \right\| . \qquad (43)$$

From (39), (41), (42), and (43), we can easily show

$$S_{x_{ij}} = \frac{|F(X)|_{x_{ij}=0}}{|F(X)|} \; \frac{\sum\limits_{m,n} |w_{mn}(x_{ij}) - w_{mn}(0)|}{\sum\limits_{m,n} |w_{mn}(x_{ij})|} \qquad (44)$$

$$\leqslant \frac{|F(X)|_{x_{ij}=0}}{|F(X)|} \left[1 + \frac{\|W\|_{x_{ij}=0}}{\|W\|} \right] . \qquad (45)$$

Finally, if we wish to study the change of $W(X)$ with respect to changes of the matrix X, we could adopt the sensitivity matrix suggested by Cruz and Perkins[5]

$$S_X = (\Delta W)\, W^{-1}\, X\, (\Delta X)^{-1} \Big|_{\Delta X \to 0} \qquad (46)$$

provided that $W(X)$ and X are square matrices of $n \times n$.

Since

$$W = D + BX(1 - CX)^{-1}A$$

$$W + \Delta W = D + B(X + \Delta X)\,(1 - CX - C\Delta X)^{-1}A$$

$$\Delta W = B\left[(X + \Delta X)\,(1 - CX - C\Delta X)^{-1} - X(1 - CX)^{-1} \right] A.$$

For $\Delta X \to 0$

$$\Delta W \approx B(1 - XC)^{-1} \; \Delta X (1 - CX)^{-1}A.$$

It is easily shown that

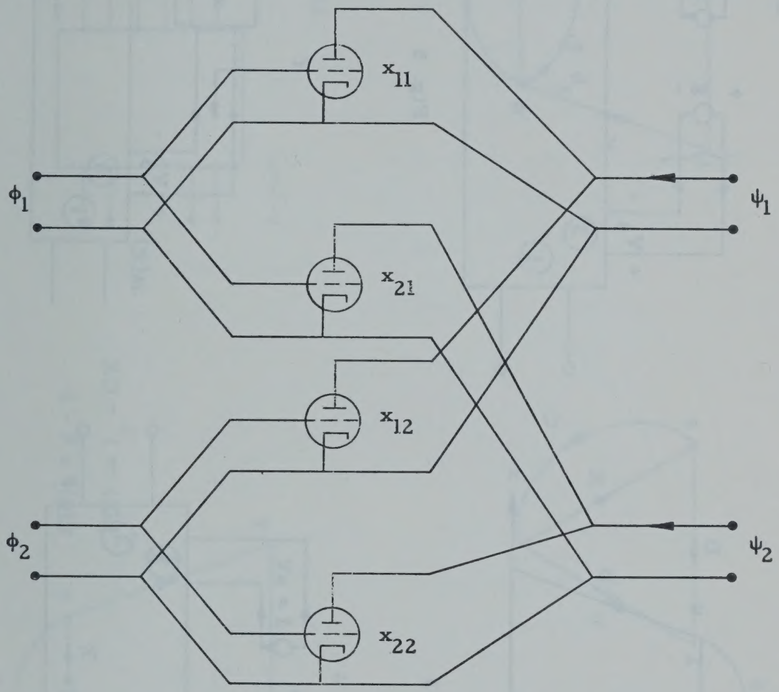
$$S_X = B(1 - XC)^{-1} \; \Delta X\, X^{-1} \; B^{-1}(1 - W(0)\, W^{-1})\; X(\Delta X)^{-1}$$

and the determinant has the simple form

$$|S_X| = |1 - XC|^{-1} \quad |1 - W(0)W^{-1}|$$

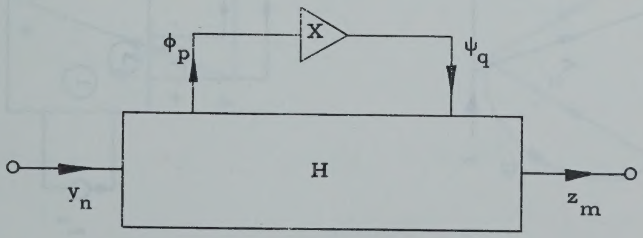
$$= \frac{1}{|F(X)|} \quad |1 - W(0)W^{-1}|. \qquad (47)$$

To summarize, we have expressed various forms of sensitivity in terms of the return difference matrix and the transmittance function similar to the classical result of Bode for the single element case.



x_{11}
x_{21}
x_{12}
x_{22}
ϕ_1
ϕ_2
ψ_1
ψ_2

Fig. 1



X
ϕ_p
ψ_q
H
y_n
z_m

Fig. 2

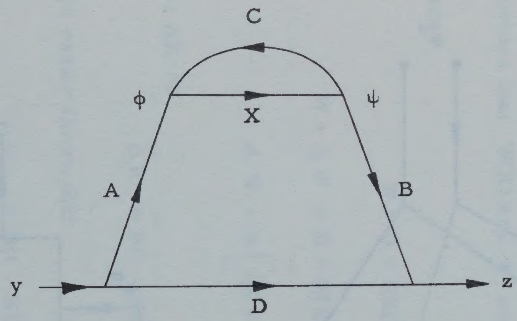
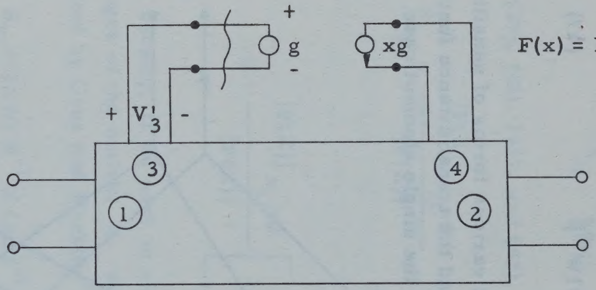


Fig. 3



$$F(x) = 1 - \frac{V_3'}{g}$$

Fig. 5

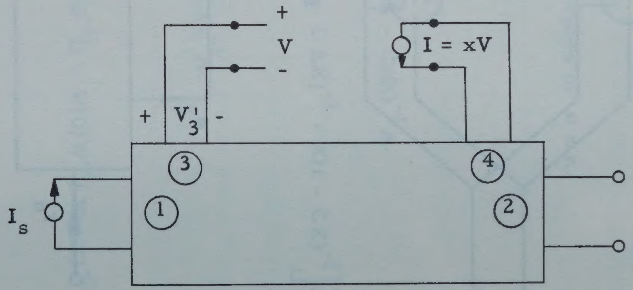
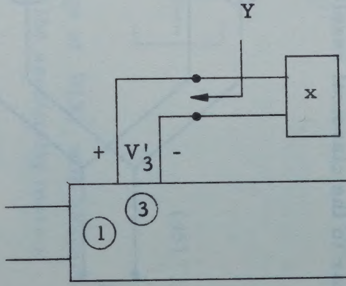
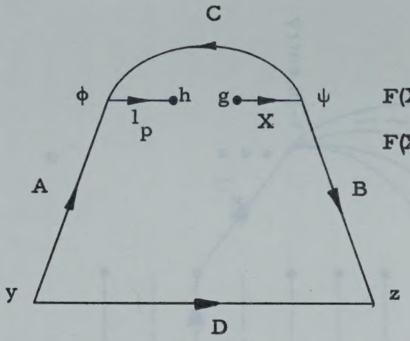

+
V
-
I = xV
+ V'3 -
3
4
Is
1
2

Fig. 4



$$F(x) = \frac{x + Y}{Y}$$

Fig. 6


C
ϕ
h
g
ψ
1_p
X
A
B
y
z
D

$F(X)g = g - h$

$F(X) = 1_p - CX$

Fig. 7

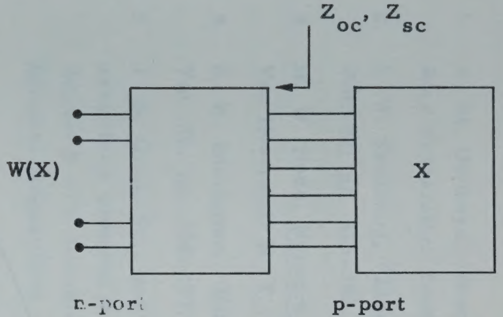

Z_{oc}, Z_{sc}
W(X)
X
n-port
p-port

Fig. 9

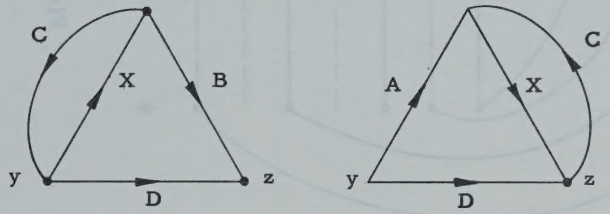

C
X
B
y
z
D
A
X
C
y
z
D

Fig. 8

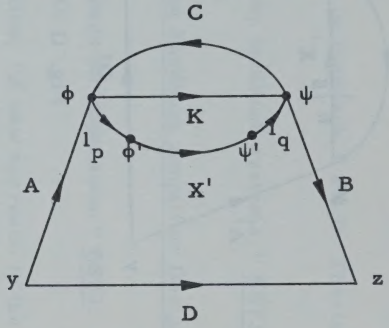

C
ϕ
ψ
K
1_p
ϕ'
ψ'
1_q
A
X'
B
y
z
D

Fig. 10

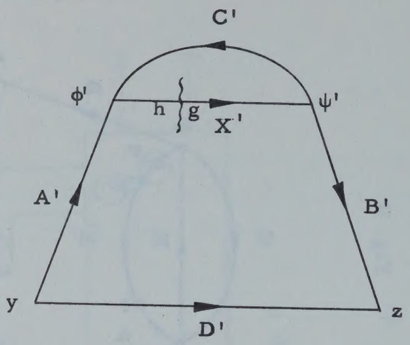
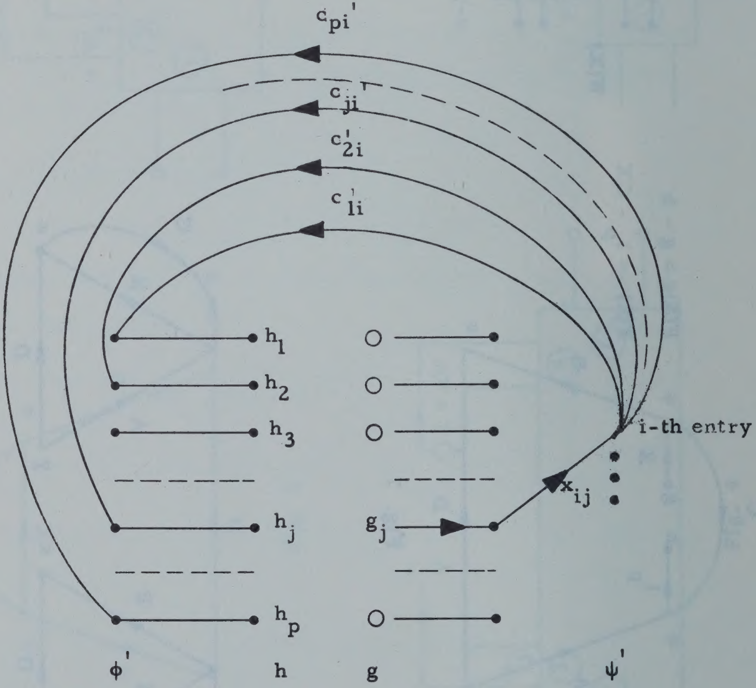

C'
φ'
ψ'
h
g
X'
A'
B'
y
z
D'

Fig. 11


c_{pi}'
c_{ji}'
c_{2i}'
c_{1i}'
h_1
h_2
h_3
h_j
g_j
h_p
x_{ij}
i-th entry
φ'
h
g
ψ'

$$F(x_{ij}) = 1 - \frac{h_j}{g_j} = 1 - c_{ji}' x_{ij}$$

Fig. 12

REFERENCES

1. R. M. Goldwyn, "Some theorems on feedback," Ph. D. Thesis, Rice University, Texas, 1960.
2. I. W. Sandberg, "Linear multi-loop feedback systems," BSTJ, Vol. 42, pp. 355-382; March 1963.
3. H. W. Bode, Network Analysis and Feedback Amplifier Design, Van Nostrand, N. Y., 1945.
4. R. B. Blackman, "Effect of Feedback on impedance," BSTJ, Vol. 22, pp. 268-277; October 1943.
5. J. B. Cruz, Jr., and W. R. Perkins, "A new approach to the sensitivity problem in multivariable feedback system design," Report R-177, Coordinated Science Laboratory, University of Illinois, August 1963.

SOME BOUNDS ON SENSITIVITY IN RLC NETWORKS*

M. L. BLOSTEIN
Department of Electrical Engineering
McGill University
Montreal, Canada

It is convenient to study the effects of parameter variations on the magnitude and phase characteristics of network functions by means of the sensitivity function first introduced by Bode.[1] The central concept here is to estimate the relative change in a network function due to a fractional change in an element value by a first order approximation in a manner amenable to both theoretical development and quantitative calculation. The approximation is sufficiently accurate for many applications and has been widely used in active networks and feedback systems generally.

The sensitivity function relates to the effects of the perturbation of a single network element. A logical extension, however, permits the effects of simultaneous variations in many elements to be treated by this concept. This extension equates the complex normalized change in a network function to a weighted sum of single parameter sensitivity functions taken with respect to the elements which are subject to change.

Because of the inextricable relationships existing between network functions, network topology, and element values, few general, quantitative statements have been formulated concerning the effects of multiparameter variations in networks. The objective of this paper is to show that multiparameter sensitivity theory can be used effectively to contribute quantitative insight into the sensitivity behavior of networks, particularly in relation to two-element kind and terminated lossless networks.

The following represents a brief outline of this paper: Section II reviews quickly the fundamental properties of the sensitivity function, stressing the concepts which favor application to RLC networks, and introduces the concept of the worst-case deviation as a figure of merit for multiparameter sensitivity. In Section III, some useful relations are established among the various sensitivity functions which can be defined in a network for a given network function. These results are applied in Sections IV and V to derive lower bounds to the worst-case deviation of a given response characteristic for a prescribed set of element tolerances which are valid for rather broad classes of networks: two-element kind and terminated lossless networks. These bounds are expressed in terms of the desired network specification and other useful design parameters.

---

*This work was supported by the Air Force Office of Scientific Research under Grant No. AF 177-63 when the author was with the Dept. of Elec. Eng., University of Illinois, Urbana, Illinois.

II. PRELIMINARIES

If a network contains elements $x_1, \ldots\ldots, x_b$ and these are subject to fractional changes $dx_1/x_1, \ldots\ldots, dx_b/x_b$, then the corresponding normalized complex change in a network function $F = F(x_i, s)$* may be estimated to a first approximation by the relation[2,3]

$$\frac{\Delta F}{F} \approx \frac{dF}{F} = \sum_{i=1}^{b} S_{x_i}^{F} \frac{dx_i}{x_i} \qquad (1)$$

where

$$S_{x_i}^{F} \triangleq \frac{x_i}{F}\frac{\partial F}{\partial x_i} = \frac{\partial \ln F}{\partial \ln x_i}$$

The quantity $S_{x_i}^{F}$ represents the sensitivity of F with respect to x and is the reciprocal of the definition originally put forward by Bode.[4] Thus $\Delta F/F$ is represented as a weighted sum of sensitivity functions, where the weights are the corresponding fractional changes in the elements. Note, since the x_i's are network elements (i.e. R, L, etc.) and therefore are real quantities, it follows from the definition of $S_{x_i}^{F}$ that,

$$\frac{\Delta\,|F(j\omega)|}{|F(j\omega)|} \approx \sum_{i=1}^{b} [\mathrm{Re}\ S_{x_i}^{F}(j\omega)]\,\frac{dx_i}{x_i} \qquad (2)$$

$$\Delta \arg F(j\omega) \approx \sum_{i=1}^{b} [\mathrm{Im}\ S_{x_i}^{F}(j\omega)]\,\frac{dx_i}{x_i} \qquad (3)$$

Hence, Equation (1) is a compact expression from which both magnitude and phases changes can be approximated readily.

An important advantage of the normalized definition of the sensitivity function described above is that $S_{x_i}^{F}$ can be characterized conveniently in terms of network determinants and cofactors. Thus, any network function may be expressed as

$$F(x_i, s) = \frac{W(x_i, s)}{T(x_i, s)}$$

where $W = W(x_i, s)$ and $T = T(x_i, s)$ represent cofactors or determinants of a system of mesh or nodal equations. Further, W and T are homogeneous polynomials in the network elements and are also linear in each x_i.[5] Hence

---

* The notation $F(x_i, s)$ is used here to indicate that F is a function of $x_1, \ldots\ldots, x_b$ as well as of s. To shorten notation even further, $F(x_i, s)$ will be denoted F(s) if it is not specifically desired to emphasize the functional dependence of F upon the elements.

$$F(x_i,s) = \frac{W^o_{x_i} + x_i W^1_{x_i}}{T^o_{x_i} + x_i T^1_{x_i}} \tag{4}$$

where $W^o_{x_i}$ and $T^o_{x_i}$ represent W and T, respectively, with $x_i = 0$, and $W^1_{x_i}$ and $T^1_{x_i}$ are minors of W and T, respectively. If the operations indicated by the definition of the sensitivity function are performed on the expression for $F(x_i,s)$ in Equation (4), the following useful formula for $S^F_{x_i}$ can be obtained:[6]

$$S^F_{x_i} = \frac{T^o_{x_i}}{T} - \frac{W^o_{x_i}}{W} \tag{5}$$

In order to discuss the multiparameter sensitivity behavior of electric networks on a quantitative basis, it is necessary to define an adequate figure of merit for network sensitivity. In this context, a useful concept is the worst-case deviation from a nominal magnitude (or phase) characteristic when a set of elements in a network have prescribed tolerances. This will be interpreted here as follows: suppose m elements in a network have values which lie in an interval I defined by

$$n_i(1 - \epsilon) \leq x_i \leq n_i (1 + \epsilon) \quad i = 1, 2, \ldots m$$

where n_i is the nominal value of element x_i and ϵ is some positive number, i.e. the tolerance on these elements is ± $100\epsilon\%$. Then, for every ω it is possible to select element values, which lie in I, such that the deviation of a given magnitude function from a nominal curve is a maximum. This maximum will be referred to as the worst-case deviation for the particular ω under consideration, and it may be formulated in terms of multiparameter sensitivity theory provided a first approximation is a reasonable estimate of the actual change. A careful examination of Equation (2) reveals that the maximum deviation will result when the x_i's possess extremal values; therefore,

$$\text{max. deviation} \approx \epsilon \, M(\omega) \tag{6}$$

where

$$M(\omega) = \sum_{i=1}^{m} \left| \operatorname{Re} S^F_{x_i} (j\omega) \right| \tag{7}$$

Figure (1) illustrates this concept graphically. For any combination of element values lying in some interval I, the magnitude characteristic, to a first approximation, must lie between the two dotted curves.

The following notation will be adopted: if the set of elements x_i refers to all resistors in a network, then $M(\omega)$ will be designated as $M_R(\omega)$. $M_L(\omega)$ and $M_C(\omega)$ will be defined in a similar manner.

III. RELATIONS AMONG SENSITIVITY FUNCTIONS IN A NETWORK

In an electric network containing b elements, b sensitivity functions can be defined for any given network function and the question arises as to whether relations exist among them. Suppose a network contains ℓ inductors, r resistors, and c capacitors, where $b = \ell + r + c$. Denote the individual elements by ℓ_i, r_i, and c_i (the reciprocal of these will be denoted by γ_i, g_i, and e_i, respectively) and consider an impedance network function $F = F(r_i, \ell_i, e_i, s)$. From the amplitude scaling property of linear networks it follows that,

$$F(ar_i, a\ell_i, ae_i, s) = aF(r_i, \ell_i, e_i, s) \tag{8}$$

where a is an arbitrary parameter. Differentiating both sides of (8) with respect to a and setting a = 1 in the resulting expression yields

$$\sum_{i=1}^{\ell} \frac{\partial F}{\partial \ell_i}\, \ell_i + \sum_{i=1}^{r} \frac{\partial F}{\partial r_i}\, r_i + \sum_{i=1}^{c} \frac{\partial F}{\partial e_i}\, e_i = F$$

hence

$$\sum_{i=1}^{\ell} S_{\ell_i}^F + \sum_{i=1}^{r} S_{r_i}^F + \sum_{i=1}^{c} S_{e_i}^F = 1 \tag{9}$$

Since the logarithmic character of $S_{x_i}^F$ implies that

$$S_{x_i}^F = - S_{1/x_i}^F \tag{10}$$

Equation (9) may be written as

$$\mathcal{S}_R^F + \mathcal{S}_L^F - \mathcal{S}_C^F = 1 \tag{11}$$

where

$$\mathcal{S}_R^F = \sum_{i=1}^{r} S_{r_i}^F, \text{ etc.}$$

Equation (11) represents a linear relation which must be satisfied by the various sensitivity functions in a network. Note that the righthand side of (11) is equal to zero or minus one when F represents a dimensionless transfer ratio or an admittance function, respectively.

A further relation among sensitivity functions can be established by employing the concept of frequency scaling. This concept is expressed by the equation

$$F(r_i, a\ell_i, ac_i, s) = F(r_i, \ell_i, c_i, as) \qquad (12)$$

where a is an arbitrary parameter and $F = F(r_i, \ell_i, c_i, s)$ is any network function. Once again, differentiating both sides of (12) with respect to a, and setting a = 1 in the resulting equation yields,

$$\sum_{i=1} \ell_i \frac{\partial F}{\partial \ell_i} = \sum_{i=1}^{c} c_i \frac{\partial F}{\partial c_i} = s\frac{dF}{ds}$$

which may be written as

$$\mathcal{S}_L^F + \mathcal{S}_C^F = \frac{d\ln F}{d\ln s} \qquad (13)$$

The latter equation possesses an interesting interpretation on the jω-axis for then,

$$\frac{d\ln F(j\omega)}{d\ln (j\omega)} = \frac{d\alpha(\omega)}{d\ln\omega} + j\frac{d\beta(\omega)}{d\ln\omega}$$

where

$$\alpha(\omega) = \ln\left|F(j\omega)\right| \triangleq \text{attenuation function}$$

$$\beta(\omega) = \arg F(j\omega) \triangleq \text{phase function}$$

Hence

$$\operatorname{Re}\mathcal{S}_L^F(j\omega) + \operatorname{Re}\mathcal{S}_C^F(j\omega) = \frac{d\alpha(\omega)}{d\ln(\omega)} \qquad (14)$$

$$\operatorname{Im}\mathcal{S}_L^F(j\omega) + \operatorname{Im}\mathcal{S}_C^F(j\omega) = \omega T_d \qquad (15)$$

where

$$T_d = \frac{d\beta(\omega)}{d\omega} \triangleq \text{group delay function.}$$

Equation (14) may be interpreted physically as follows: in any network, if all inductances and capacitances undergo equal normalized perturbations, the resulting change in the magnitude characteristic of any network function is independent of the realization technique used to obtain the network, and depends only on the slope of the prescribed attenuation curve. A similar result follows for the phase characteristic from Equation (15).

Equations (11) and (13) represent fundamental relations between sensitivity functions and are valid for any RLC network. However the restriction to RLC elements is not essential to the proof of these relations. Equation (11) results because $F(\ell_i, r_i, e_i, s)$ is a homogeneous function of degree one in the elements and the derivation parallels the proof of Euler's Theorem on homogeneous functions.[7] If active elements are included, network functions are still homogeneous in the elements provided the active elements can be represented by an equivalent circuit containing transfer immittances, or can be converted into such a form in conjunction with the remainder of the network. Thus an equation analogous to (11) can be obtained for the active case. It is fairly obvious that Equation (13) is valid for both active and passive networks.

IV. APPLICATION TO TERMINATED LOSSLESS NETWORKS

The theory developed in the last section will be applied here to establish lower bounds to $M_C(\omega)$ and $M_L(\omega)$ for transfer functions of terminated LC networks. The development will assume a doubly-terminated LC structure in which the network function of interest is the voltage transfer ratio. The extension of these results to other cases in this network class is a relatively simple matter and will not be discussed here.

Consider the network structure shown in Figure 2 and suppose the network function of interest is $V = v_o/v_i$. Under the conditions stipulated, Equations (11) and (13) become

$$\mathcal{S}_R^V + \mathcal{S}_L^V - \mathcal{S}_C^V = 0$$

$$\mathcal{S}_L^V + \mathcal{S}_C^V = \frac{d\ln V}{d\ln s}$$

which may be solved to yield

$$\mathcal{S}_L^V = \frac{1}{2}\left[\frac{d\ln V}{d\ln s} - \mathcal{S}_R^V\right] \tag{16}$$

$$\mathcal{S}_C^V = \frac{1}{2}\left[\frac{d\ln V}{d\ln s} + \mathcal{S}_R^V\right] \tag{17}$$

For the structure shown in Figure 2, $\mathcal{S}_R^V$ may be expressed in terms of the scattering parameters s_{ij} of the lossless network normalized with respect to the terminating resistors (See Appendix 1). The result is given by

$$\mathcal{S}_R^V = \frac{1}{2}\left(s_{11} + s_{22}\right) \qquad (18)$$

Let $s_{ij}(j\omega) = \left|s_{ij}\right| \exp\left(j\phi_{ij}\right)$*; then, due to the lossless, reciprocal nature of the two port network, the scattering parameters are constrained by the relations,[8]

$$s_{12} = s_{21}$$

$$\left|s_{11}\right| = \left|s_{22}\right|$$

$$\phi_{22} = 2\phi_{12} - \phi_{11} + \pi$$

As a result, Equation (18) may be written as

$$\mathcal{S}_R^V (j\omega) = \frac{1}{2}\left|s_{11}\right|\left[\exp(j\phi_{11}) - \exp(j2\phi_{12} - j\phi_{11})\right]$$

so that

$$\mathrm{Re}\, \mathcal{S}_R^V (j\omega) = \left|s_{11}\right|\left[\sin\phi_{12}\, \sin(\phi_{12} - \phi_{11})\right] \qquad (19)$$

whence

$$\left|\mathrm{Re}\, \mathcal{S}_L^V (j\omega)\right| = \frac{1}{2}\left|\frac{d\alpha(\omega)}{d\ln\omega} - \left|s_{11}\right|\left[\sin\phi_{12}\, \sin(\phi_{12} - \phi_{11})\right]\right| \qquad (20)$$

$$\left|\mathrm{Re}\, \mathcal{S}_C^V (j\omega)\right| = \frac{1}{2}\left|\frac{d\alpha(\omega)}{d\ln\omega} + \left|s_{11}\right|\left[\sin\phi_{12}\, \sin(\phi_{12} - \phi_{11})\right]\right| \qquad (21)$$

Since

$$\sum_{i=1}^{m}\left|\mathrm{Re}\, S_{x_i}^F (j\omega)\right| \geqslant \left|\sum_{i=1}^{m} \mathrm{Re}\, S_{x_i}^F (j\omega)\right| \qquad (22)$$

it follows that

$$M_L(\omega) \geqslant \left|\mathrm{Re}\, \mathcal{S}_L^V (j\omega)\right| \qquad (23)$$

---

*It should be kept in mind that $\left|S_{ij}\right|$ and ϕ_{ij} are functions of ω. This has not been shown explicitly in order to shorten notation.

$$M_C(\omega) \geqslant \left| \mathrm{Re}\, \mathcal{S}_C^V (j\omega) \right| \qquad (24)$$

Hence, Equations (20) and (21) represent lower bounds to $M_L(\omega)$ and $M_C(\omega)$ which are valid for every ω.

The significance of the various terms in (20) and (21) are more apparent when considering the design of doubly-terminated networks by insertion loss techniques, for then $|s_{12}|$ and $|s_{11}|$ are quantities determined by the network specification $\alpha(\omega)$; further, if a minimum phase realization is desired, then ϕ_{12} is also fixed, i.e. the magnitude of a minimum phase-function uniquely determines the associated phase angle. Thus, the only parameter not fixed by the specification is ϕ_{11}, the phase angle of s_{11}. This angle is a function of which zeros of $\left[s_{11}(s) \quad s_{11}(-s) \right]$ are selected as zeros for s_{11}. In general this selection is somewhat arbitrary, although in certain circumstances it may be dictated by practical considerations.[9]

The relative magnitudes of the terms in (20) and (21) should also be noted; where the slope of the attenuation curve is large (viewed on a logarithmic frequency scale), such as in the transition band of a filter, the first term is dominant, particularly since the passive nature of the network restricts $|s_{11}| \leq 1$. Conversely, in regions of the frequency spectrum where the attenuation curve remains flat, the second term will be more significant. Numerically, if the slope of attenuation curve is 6 db/octave, then $d\alpha(\omega)/d\ln\omega = 1$ so that a filter which has a cut-off rate of 60 db/octave in the transition band has $M_C > 4$ there regardless of how the network is designed.

Finally, it is clear that the lower bounds to $M_C(\omega)$ and $M_L(\omega)$ are the same for all network realizations which have the same two-port parameters for the LC network. Thus, all equivalent networks obtained from a known realization by means of Cauer's linear transformation techniques[10] have the same lower bound to $M_C(\omega)$ and $M_L(\omega)$.

V. APPLICATION TO RC NETWORKS

When a network contains only resistive and capacitive elements, Equations (11) and (13) can be solved explicitly for $\mathcal{S}_C^F$ and $\mathcal{S}_R^F$. If F(s) represents an impedance function then

$$\mathcal{S}_R^F = 1 + \frac{d\ln F(s)}{d\ln s} \qquad (25)$$

$$\mathcal{S}_C^F = \frac{d\ln F(s)}{d\ln s} \qquad (26)$$

Therefore

$$\mathrm{Re}\ \mathscr{S}_{R}^{F}(j\omega) = 1 + \frac{d\alpha(\omega)}{d\ln\omega} \tag{27}$$

$$\mathrm{Re}\ \mathscr{S}_{C}^{F}(j\omega) = \frac{d\alpha(\omega)}{d\ln\omega} \tag{28}$$

Equations (26) and (27) together with Equation (22) show that

$$M_R(\omega) \geq \left| 1 + \frac{d\alpha(\omega)}{d\ln\omega} \right| \tag{29}$$

$$M_C(\omega) \geq \left| \frac{d\alpha(\omega)}{d\ln\omega} \right| \tag{30}$$

The latter inequalities are interesting since they place lower bounds on M_R and M_C which are valid at every ω and for any RC network. In fact, the bounds depend solely on $\alpha(\omega)$ and hence represent the minimum theoretical values of M_R and M_C for the desired network specification.

The question arises as to whether a network can be designed to achieve these minimum values. It is clear that a necessary and sufficient condition for Equations (29) and (30) to be equalities is that all sensitivity functions with respect to elements of one kind (i.e. resistors or capacitors) must have real parts with the same sign. It is difficult to determine whether it is possible to satisfy this condition, particularly if the equalities are desired only over a limited range of ω, although examples can be constructed easily where the condition must be violated at any ω.

An example for which the equalities are assured at every ω is the simple low-pass ladder shown in Figure 4 where F(s) is assumed to be the transfer impedance from 11' to 22'. This result can be established by means of the following two theorems which are derived in Appendix 2.

Theorem 1: In any RC network, $T^o_{g_i}/T$ has the properties of an RC driving point admittance function; hence $\mathrm{Re}\left[T^o_{g_i}/T\,(j\omega)\right]$ is a positive monotone increasing function of ω which is bounded above by unity. $W^o_{g_i}/W$ possesses similar properties if W is a determinant of an RC network.

Theorem 2: In any RC network, $T^o_{c_i}/T$ has the properties of an RC driving point impedance function; hence $\mathrm{Re}\left[T^o_{c_i}/T\,(j\omega)\right]$ is a positive monotone decreasing function of ω bounded above by unity. $W^o_{c_i}/W$ possesses similar properties if

W is a determinant of an RC network.

For the network in Figure 4 it follows from Equations (5) and (10) that

$$\text{Re } S^F_{r_i}(j\omega) = -\text{Re } \frac{T^o_{g_i}}{T}(j\omega)$$

$$\text{Re } S^F_{c_i}(j\omega) = \text{Re } \frac{T^o_{c_i}}{T}(j\omega) - 1$$

Clearly both Re $S^F_{r_i}(j\omega)$ and Re $S^F_{c_i}(j\omega)$ are negative at every ω. Therefore,

$$M_R(\omega) = \left| 1 + \frac{d\alpha(\omega)}{d\ln\omega} \right|$$

$$M_C(\omega) = \left| \frac{d\alpha(\omega)}{d\ln\omega} \right|$$

Similar results follow for the simple high-pass ladder and when F(s) is any transfer immittance or transfer ratio. In the synthesis of an all-pole Z_{12}, it is required to choose a realizable Z_{11} with the same poles as Z_{12} and the synthesis proceeds by means of a Cauer development of this driving-point impedance;[10] the above discussion shows that the worst-case deviation is invariant to the choice of Z_{11}.

VI. CONCLUSION

The lower bounds to the M functions computed here merit attention since they represent quantitative statements about network sensitivity which are valid for broad classes of networks. For these classes, they provide a measure of the intrinsic sensitivity associated with the realization of a given network specification and demonstrate the intimate relationship existing between network sensitivity and network functions.

APPENDIX 1.

From Equations (4) and (10), $S^V_{R_s}$ in Figure 2 is given by

$$S^V_{R_s} = -\frac{T^o_{G_s}}{T}$$

$$= -\frac{T^o_{G_s}}{T^o_{G_s} + G_s T'_{G_s}}$$

$$= -\frac{1}{1 + \frac{Z_i}{R_s}}$$

where Z_i is defined by Figure 3. But,

$$Z_i/R_s = \frac{1 + s_{11}}{1 - s_{11}}$$

so that

$$S^V_{R_s} = - \frac{1}{2} (1 - s_{11})$$

Further,

$$S^V_{R_L} = - (\frac{T^o_{G_L}}{T} - 1)$$

hence an analysis similar to the above yields,

$$S^V_{R_L} = 1 - \frac{1}{2} (1 - s_{22})$$

Therefore,

$$\mathcal{S}^V_R = S^V_{R_s} + S^V_{R_L} = \frac{1}{2} (s_{11} + s_{22})$$

APPENDIX 2.

The proof of Theorem 1 is based on the observation that $T^o_{g_i}$ is the nodal determinant of the RC network with element g_i removed (open-circuited) and T'_{g_i} represents the network determinant when element g_i is shorted out. Thus $T'_{g_i}/T^o_{g_i}$ is the driving point impedance which element g_i faces. Let this impedance be denoted by Z_{g_i}. Since,

$$\frac{T^o_{g_i}}{T} = \frac{T^o_{g_i}}{T^o_{g_i} + g_i T'_{g_i}}$$

$$= \frac{1/g_i}{1/g_i + T'_{g_i}/T^o_{g_i}}$$

$$= \frac{r_i}{r_i + Z_{g_i}}$$

then, the function $T^o_{g_i}/T$ has all the properties of an RC driving point admittance function; in particular Re $\left[T^o_{g_i}/T\ (j\omega)\right]$ is a positive monotone increasing function of ω.

In order to show that Re $\left[T^o_{g_i}/T\ (j\omega)\right]$ is bounded above by unity, write

$$Z_{g_i}(j\omega) = R_{g_i}(\omega) + j\,X_{g_i}(\omega)$$

so that

$$\mathrm{Re}\left[T^o_{g_i}/T\ (j\omega)\right] = \frac{r_i\left[r_i + R_{g_i}(\omega)\right]}{\left[r_i + R_{g_i}(\omega)\right]^2 + \left[X_{g_i}(\omega)\right]^2}$$

$$= \frac{1 + R_{g_i}(\omega)/r_i}{\left[1 + R_{g_i}(\omega)/r_i\right]^2 + \left[X_{g_i}(\omega)/r_i\right]^2}$$

Since $R_{g_i}(\omega) > 0$, it is obvious that

$$\mathrm{Re}\left[T^o_{g_i}/T\ (j\omega)\right] \leq 1$$

An identical analysis may be applied to $W^o_{g_i}/W$ provided $W'_{g_i}/W^o_{g_i}$ represents an RC driving point impedance. This is always the case if W is a determinant of an RC network.

The proof of Theorem 2 parallels that of Theorem 1. Here,

$$\frac{T^o_{c_i}}{T} = \frac{\dfrac{1}{sc_i}}{\dfrac{1}{sc_i} + Z_{c_i}}$$

$$= \frac{1/c_i}{1/c_i + sZ_{c_i}}$$

where Z_{c_i} is the impedance which element c_i faces. Since Z_{c_i} is an RC driving point impedance, sZ_{c_i} is an RC driving point admittance and $T^o_{c_i}/T$ has the properties of an RC driving point impedance function. Therefore, Re $\left[T^o_{c_i}/T\ (j\omega)\right]$ is a positive monotone decreasing function of ω. The remainder of the proof is identical to that of Theorem 1.

REFERENCES

1. H.W. Bode, "Network Analysis and Feedback Amplifier Design", D. Van Nostrand Co., Inc., Princeton, N. J., 1945; p. 52.

2. J.J. Mikulski, "A Correlation between Classical and Pole-Zero Sensitivity", Technical Note No. 5, Contract No. AF49(638)-63 Circuit Theory Group, Electrical Engineering Research Labs., University of Illinois; February 1959.

3. A.J. Goldstein and F.F. Kuo, "Multiparameter Sensitivity", IRE Transactions on Circuit Theory, Vol. CT-8, pp. 177-8; June, 1961.

4. J.G. Truxal, "Control System Synthesis", McGraw-Hill Book Co., Inc., New York, N.Y., 1955; p. 120.

5. N.F. Tsang, "On Electrical Network Determinants", J. Math. and Phys., Vol. 33, pp. 185-193; July 1954.

6. W.A. Lynch, "A Formulation of the Sensitivity Function", IRE Transactions of Circuit Theory, Vol. CT-4, p. 289; September 1957.

7. W. Kaplan, "Advanced Calculus", Addison-Wesley Publishing Co., Inc., Cambridge, Mass., 1953; p. 90.

8. H.J. Carlin, "The Scattering Matrix in Network Theory", IRE Transactions on Circuit Theory, Vol. CT-3, pp. 88-97; June 1956.

9. L. Weinberg, "Network Analysis and Synthesis", McGraw-Hill Book Co., New York, N. Y., 1962.

10. E.A. Guillemin, "Synthesis of Passive Networks", John Wiley and Sons, Inc., New York, N. Y., 1957.

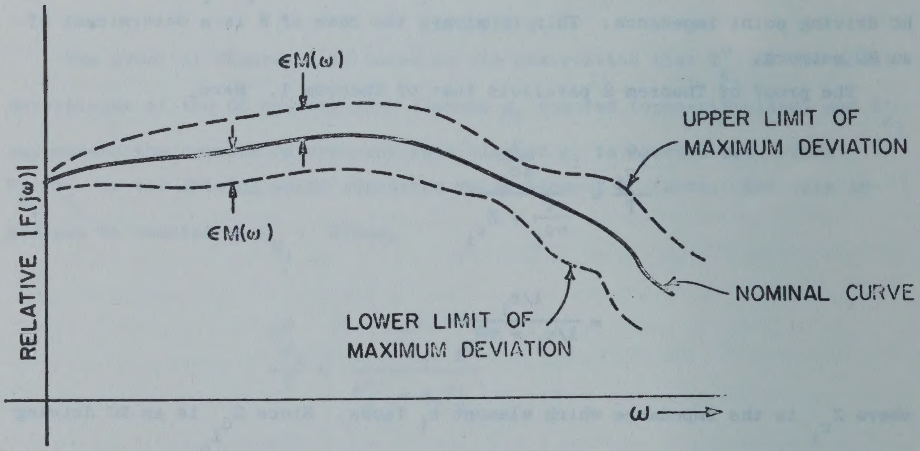


Figure 1 - Graphical Illustration of Worst-Case Deviation

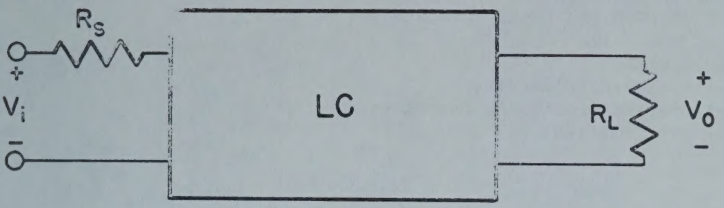

R_S
+
V_i
−
LC
R_L
+
V_o
−

Figure 2 - A Lossless Doubly-Terminated Network

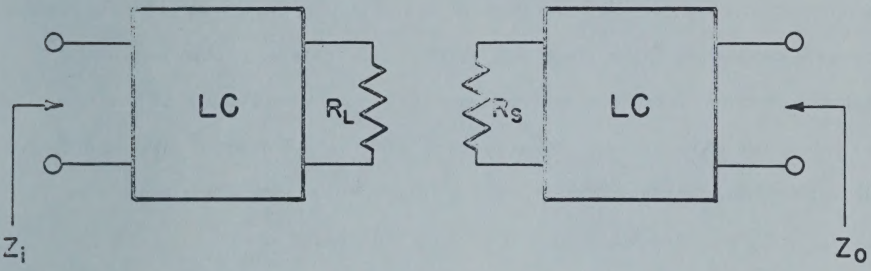

LC
R_L
R_S
LC
Z_i
Z_o

Figure 3 - Illustration of Z_i and Z_o

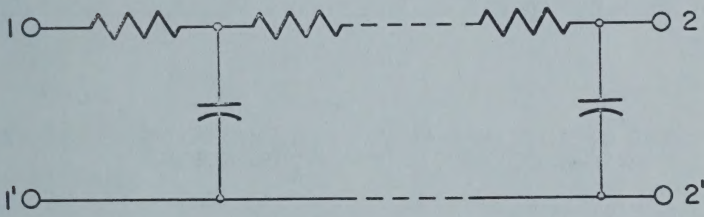


Figure 4 - A simple Low-Pass RC Ladder

ON THE UPPER BOUND AND LOWER BOUND OF EXCURSIONS OF CRITICAL-FREQUENCIES OF A PERTURBED NETWORK*

M. J. REMEC
Northwestern University
Information Processing Laboratory
Evanston, Illinois

INTRODUCTION

The pole (or zero) sensitivity for a multiple-pole (or a multiple zero), if the definition is forced, is infinite. As such, the pole (or zero) sensitivity is useless as an indication of the relative change of the pole (or zero) location resulting from a perturbation (deviation from nominal values) of the system-function parameters. As an example of perturbation of the parameters of the system function the case of circuit element deviations due to inaccurate element values, changing environment or aging, may be cited when one is dealing with networks.

In some problems the largest and the smallest change of either a multiple or a simple pole (or zero) is of interest. For small perturbation, an upper bound to the root movements can be found via root sensitivity if the root is simple. The bounds found in this paper[3] are free of this restriction.

UPPER BOUND ON THE ZERO MOVEMENTS BASED ON THE POLYNOMIAL COEFFICIENT PERTURBATION

Two theorems[2] to be used are:

1.) The zeros of a polynomial are continuous functions of the coefficients of the polynomial.

2.) The Szegö convolution theorem which reads: If from the given polynomials

$$A(s) = \sum_{i=0}^{n} \binom{n}{i} a_i s^i \qquad (1)$$

$$= a_n \prod_{i=1}^{n} (s-z_i) \qquad (2)$$

---

*The financial support of the Air Force Office of Scientific Research, Air Force Research and Developement Command, for this work is gratefully acknowledged.

$$B(s) = \sum_{i=0}^{n} \binom{n}{i} b_i s^i \tag{3}$$

a third polynomial

$$C(s) = \sum_{i=0}^{n} \binom{n}{i} c_i s^i \tag{4}$$

$$= \sum_{i=0}^{n} \binom{n}{i} a_i b_i s^i$$

is formed, and if all the zeros β_j of B(s) lie in an arbitrary convex region B, then every zero of C(s) has the form

$$\gamma = - \alpha \beta \tag{5}$$

where β is a suitably chosen point in B and α is a zero of A(s).

In perturbed polynomials, such as the $N(s, \overline{x})$ or $D(s, \overline{x})$ of the system function whose coefficients are altered slightly because of parameter changes,

$$\overline{\Theta} = (\Theta_{n-1}, \Theta_{n-2}, \ldots, \Theta_0) \tag{6}$$

gives the percentage change in every coefficient but a_n, which is held fixed. Note that this is not a restriction because if a_n is perturbed its effect may be accounted for by dividing the coefficients a_i by $(1 + \Theta_n)$ and calculating the correspondent $\overline{\Theta}$. The perturbed polynomial thus has the form,

$$\tilde{A}(s) = a_n \binom{n}{n} s^n + a_{n-1} (1+\Theta_{n-1})\binom{n}{n-1} s^{n-1} + \ldots + a_0(1+\Theta_0)\binom{n}{0} s^0 \tag{7}$$

having zeros

$$\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n, \tag{8}$$

where "$\sim$" denotes the quantities resulting from perturbation.

If the polynomial in (8)

$$B(s) = B(s, \overline{\Theta}) \tag{9}$$

is formed in the fashion

$$B(s, \overline{\Theta}) = (s + 1)^n + \sum_{i=0}^{n-1} \Theta_i \binom{n}{i} s^i \tag{10}$$

$$= \sum_{i=0}^{n} \binom{n}{i} s^i + \sum_{i=0}^{n-1} \Theta_i \binom{n}{i} s^i \tag{11}$$

$$= \sum_{i=0}^{n} (1+\Theta_i)\binom{n}{i} s^i - \Theta_n s^n \tag{12}$$

then the convolved polynomial $C(s, \overline{\Theta})$ is

$$C(s, \overline{\Theta}) = A(s) * B(s, \overline{\Theta}) \tag{13}$$

$$= [\, a_n\binom{n}{0}s^n + a_{n-1}\binom{n}{i}s^{n-1} + \ldots + a_0\binom{n}{0}s^0] *[\sum_{i=0}^{n} +(1+\Theta_i)\binom{n}{i}s^i - \Theta_n s^n] = \tag{14}$$

$$= a_n\binom{n}{n}s^n + a_{n-1}(1+\Theta_{n-1})\binom{n}{n-1}s^{n-1} + \ldots + a_0(1+\Theta_o)s^0 = \tag{15}$$

$$= \widetilde{A}(s) \ . \tag{16}$$

Thus the perturbed polynomial $\widetilde{A}(s)$ can be formed from the nominal polynomial $A(s)$ and the properly formed (10) perturbation polynomial $B(s, \overline{\Theta})$. Applying (5) it follows that every zero of $\widetilde{A}(s)$ has the form

$$\tilde{z}_i = - \beta z_j \tag{17}$$

and β is a suitably chosen point in the convex region B. Referring to zeros of $B(s, \overline{\Theta})$ as β_k, then

$$B(s, \overline{\Theta}) = \prod_{k=1}^{n} (s - \beta_k(\overline{\Theta})) \ . \tag{18}$$

In view of (10) a convenient region B, satisfying conditions of the Szegö convolution theorem is that region consisting of the interior of the circle and its boundary, centered at

$$s = -1 + j0 \tag{19}$$

and having as its radius

$$\sigma(\overline{\Theta}) = \max_{1 \le k \le n} |\beta_k + 1| ; \tag{20}$$

or, equivalently, region B is given by B:

$$|s + 1| \le \max_{1 \le k \le n} |\beta_k + 1| \ . \tag{21}$$

For any point of B, then,

$$|\beta + 1| \le \sigma(\overline{\Theta}) \ , \tag{22}$$

(22) with (17) substituted becomes

$$\left| \frac{\tilde{z}_i - z_j}{z_j} \right| \leq \sigma\,(\bar{\Theta}) \,. \tag{23}$$

Invoking the theorem on continuity of the zeros as functions of the coefficients in (10), it follows that for some $\bar{\Theta}_c$ such that

$$0 < \|\bar{\Theta}\| \leq \|\bar{\Theta}_c\| \tag{24}$$

(23) necessarily becomes

$$\left| \frac{\tilde{z}_j - z_j}{z_j} \right| \leq \sigma\,(\bar{\Theta}) \,. \tag{25}$$

If it did not, (23) would be violated. That is impossible because in view of the zero continuity mentioned,

$$\lim_{\|\bar{\Theta}\| \to 0} \sigma\,(\bar{\Theta}) = 0 \tag{26}$$

The significance of this result stems from the fact that Equation (25) has the form

$$\left| \frac{\Delta z_j}{z_j} \right| \leq \sigma\,(\bar{\Theta}) \,, \tag{27}$$

which places an u.b. on the j^{th} zero movement resulting from a simultaneous coefficient perturbation given by $\bar{\Theta}$, provided $\|\bar{\Theta}\|$ is sufficiently small so that there is no non-coincident intersection of the bounds between the adjacent zeros. This problem, if only one of the coefficients is perturbed, is the same as that treated by Calahan[1].

LOWER BOUND ON THE ZERO MOVEMENT BASED ON THE POLYNOMIAL COEFFICIENT PERTURBATION

Let z_1 be the zero for which the circular region C_1 is sought outside of which z_1 moves under coefficient perturbation $\bar{\Theta}$. Let the nominal polynomial $P(s, \bar{x}_o)$ be

$$P(s, \bar{x}_o) = (s - z_1) \sum_{i=0}^{n-1} a_i s^i \tag{28}$$

$$= \sum_{i=0}^{n} b_i s^i \tag{29}$$

and

$$Q(s) = F(s, \bar{x}) - P(s, \bar{x}_o) \tag{30}$$

where $F(s, \bar{x})$ is perturbed polynomial $P(s, \bar{x})$. Furthermore,

$$Q(s) = \sum_{i=0}^{n} \Theta_i b_i s^i \quad . \tag{31}$$

Let C_1 be centered at z_1 and have a radius ϵ. Thus C_1 is given by

$$s = z_1 + \epsilon e^{j\Theta} \quad . \tag{32}$$

Rouche's Theorem[2] states: If $P(s)$ and $Q(s)$ are analytic interior to a simple closed Jordan curve C and if they are continuous on C and

$$|P(s)| < |Q(s)| \text{ on } C, \tag{33}$$

then the function $F(s) = P(s) + Q(s)$, has the same number of zeros interior to C as does $Q(s)$.

Assuming that $Q(s)$ has no zeros interior to C_1 and applying Rouche's theorem on the boundary C_1 one has

$$|\epsilon e^{j\Theta} \sum_{i=0}^{n-1} a_i (z_1 + \epsilon e^{j\Theta})^i| < |\sum_{i=0}^{n} \Theta_i b_i (z_i + \epsilon e^{j\Theta})^i| . \tag{34}$$

Strengthening the inequality (34) by taking at least the max of LHS of (34) and at most the min of RHS of (34), to ensure that (34) holds for all angles Θ (this decreases the radius ϵ), one has

$$\epsilon \left| \left(\sum_{i=0}^{n-1} a_i \left(|z_i| + \epsilon \right)^i \right) \right| < \min_{\Theta} \left| \sum_{i=0}^{n} \Theta_i b_i \left(z_1 + \epsilon e^{j\Theta} \right)^i \right| . \qquad (35)$$

Thus the largest ϵ , call it ϵ_M, such that the inequaltiy

$$\epsilon \left| \left[\sum_{i=0}^{n-1} a_i \left(|z_i| + \epsilon \right)^i \right] \right| - \min_{\Theta} \left| \sum_{i=0}^{n} \Theta_i b_i \left(z_1 + \epsilon e^{j\Theta} \right)^i \right| < 0, \qquad (36)$$

holds for all

$$0 < \epsilon \leq \epsilon_M \qquad (37)$$

can be used as radius of C_1. The region C_1 is the region outside of which the zeros of the polynomial $P(s, \overline{x})$ will surely move when the coefficients of $P(s, \overline{x}_o)$ are perturbed by an amount $\overline{\Theta}$. Either the theorem on the smallest modulus of a zero, or a numerical scheme when a larger ϵ is desired, can be resorted to in the search of the ϵ satisfying (36).

DETERMINATION OF C_1 FOR A PARTICULAR POLYNOMIAL

To demonstrate the procedure and closeness of the approximation attained by the above method, the following perturbed polynomial is solved for the region C_3.

Let

$$P(s, \overline{x}_o) = s^3 + 4s^2 + 6s + 4 \quad , \qquad (38)$$

$$\overline{\Theta} = (\Theta_3, \Theta_2, \Theta_1, \Theta_0)$$

$$= (0, .05, .067, .1) \quad , \qquad (39)$$

so that

$$Q(s) = .2s^2 + .4s + .4 \quad . \qquad (40)$$

If the movement of the zero, z_3, at

$$s = -2 \tag{41}$$

is of interest the inequality to be satisfied on C_3 is

$$|2\epsilon e^{j\Theta} - 2\epsilon^2 e^{j2\Theta} + \epsilon^3 e^{j3\Theta}| < |.4 - .4\epsilon e^{j\Theta} + .2\epsilon e^{j2\Theta}|. \tag{42}$$

After strengthening the inequality (42) so that it surely holds for all angles Θ, (42) holds if

$$(2\epsilon + 2\epsilon^2 + \epsilon^3) < (.4 - .4\epsilon - .4\epsilon^2), \tag{43}$$

$$2.4\epsilon + 2.4\epsilon^2 + \epsilon^3 < .4 . \tag{44}$$

By trial and error it is quickly established that

$$\epsilon_M = .14 \tag{45}$$

Thus the zero at $s = -2$ moves by more that .14 units away from its nominal position under perturbation $\overline{\Theta}$, which compares reasonably with the actual zero movement of .2 to the new position at $s = -2.2$.

REFERENCES

1.) Calahan, D. A., "On the Sensitivity of the Zeros to Changes in a Coefficient," Unpublished Work of Contract AF(638)-63, University of Illinois, 1961.

2.) Marden, M., The Geometry of the Zeros of a Polynimial in a Complex Variable, American Mathematical Society, New York, pp. 47-51, 1949.

3.) Remec, M. J., "Application of Multiparameter Sensitivity in Linear Systems", Ph. D. EE Thesis, University of Illinois, Ch. 5, 1963.

THE SYNTHESIS OF SINGLE VARIABLE SYSTEMS BY STATE VARIABLE FEEDBACK

MAJOR BERNARD S. MORGAN, JR.
Air Force Office of Scientific Research
Washington, D. C.

The synthesis of a q^{th} order system with one input and one output by state variable feedback is treated. A basic assumption is that all the state variables are available and measurable. It is shown that one is able by state variable feedback to shift the poles of a system to any desired position without increasing the order of the system. The only requirement for this is that the system be controllable. Sensitivity functions are obtained and result from a new expression for the first approximation to a characteristic root. By using the sensitivity functions, procedures are obtained and are presented for selecting the poles of the closed loop system so that the effects of the changes in the plant's parameters are reduced.

I DESIGN FOR SPECIFIED CHARACTERISTIC ROOTS

A simple systematic procedure for the design of a system for specified characteristic roots is presented in this section. Design is accomplished by means of state variable feedback. A basic assumption is that all the state variables are available and measurable. The procedure for the design of a system for specified characteristic roots is applicable to the problem of Povejsil and Fuchs (1) who presented a trial and error method for the modification of the characteristic equation of a system. Bass (2) has obtained a similar design procedure by a relatively sophisticated application of matrix theory. In addition this procedure can be inferred from the recent work of Kalman (3).

Assume that a plant is described by a system of equations of the form

$$\dot{\vec{y}} = \alpha \vec{y} + \vec{\beta} u, \tag{1}$$

$$v = y_1, \tag{2}$$

where $\vec{y}$ is a q dimensional state vector with components $y_1, y_2, \ldots y_q$; α is a constant q^{th} order differential transition matrix, $\vec{\beta}$ is a constant q row, single column input matrix, u is the input, and v is the output.

It is desired that a linear closed loop system be designed so that it has specified characteristic roots. However before this is done for the plant described by equations (1) and (2) consider a similar design problem for a plant with an input-state relationship described by

$$\dot{\vec{z}} = K\vec{z} + \vec{\xi} u, \tag{3}$$

where $\vec{z}$ is a q dimensional state vector and

$$K = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ k_q & k_{q-1} & \dots & & k_1 \end{bmatrix} \quad ; \quad \vec{\xi} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \tag{4}$$

The system of (3) has a characteristic polynomial

$$f(\lambda) = \lambda^q - k_1 \lambda^{q-1} - \dots - k_q . \tag{5}$$

Compensate (3) by state variable feedback so that

$$u = \underline{n}\,\vec{z} + c_1 w, \tag{6}$$

where $\underline{n}$ is a constant feedback row vector with components n_1, n_2, ... n_q; c_1 is a constant ; w is the controlling input. Thus

$$\dot{\vec{z}} = (K + \vec{\xi}\,\underline{n})\,\vec{z} + \vec{\xi}\, c_1 w, \tag{7}$$

and so

$$\dot{\vec{z}} = H\,\vec{z} + \vec{\xi}\, c_1 w \tag{8}$$

with

$$H = \begin{bmatrix} 0 & 1 & 0 & . & . & . & 0 \\ 0 & 0 & 1 & . & . & . & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ h_q & h_{q-1} & & . & . & . & h_1 \end{bmatrix} \tag{9}$$

and characteristic polynomial

$$g(\lambda) = \lambda^q - h_1 \lambda^{q-1} - \, . \, . \, . - h_q . \tag{10}$$

From equations (7) and (8)

$$n_j = h_{q-j+1} - k_{q-j+1} , \quad j=1,2\dots,q \tag{11}$$

From eq. (11) it is seen that if the k's and h's are known one can readily solve for the n's. This suggests a procedure for changing the characteristic polynomial of a system described by equations (3) that is, if the k's, the coefficients of the open loop characteristic polynomial, are given one can easily determine the feedback gains, the n's, for a prescribed closed loop polynomial with coefficents $h_1, \ldots h_q$.

However the real aim is to compensate the system described by equation (1) and change its characteristic polynomial. This can be done by choosing the following feedback law

$$u = \underbracket{m}\,\vec{y} + c_1\ w\ , \tag{12}$$

and changing coordinates by

$$\vec{y} = F\ \vec{z}\ , \tag{13}$$

where the determinant of F is nonzero. The basis for the choice of c_1 is discussed later. Now placing equation (12) in (1) one obtains

$$\dot{\vec{y}} = (\alpha + \vec{\beta}\ \underbracket{m})\ \vec{y} + \vec{\beta}\ c_1\ w, \tag{14}$$

or

$$\dot{\vec{y}} = A\ \vec{y} + \vec{\beta}\ c_1\ w. \tag{15}$$

Using equation (13) in (14) gives

$$\dot{\vec{z}} = (F^{-1}\ \alpha\ F + F^{-1}\vec{\beta}\ \underbracket{m}\ F)\vec{z} + F^{-1}\vec{\beta}\ c_1\ w. \tag{16}$$

Comparing (16) and (7)

$$F^{-1}\ \alpha\ F = K\ ,\ F^{-1}\ \vec{\beta} = \vec{\xi},\ \underbracket{n} = \underbracket{m}\,F. \tag{17}$$

If $F = (\vec{f}_1, \vec{f}_2, \ldots \vec{f}_q)$ one obtains from (17) the following

$$\vec{f}_q = \vec{\beta}, \tag{18}$$

$$\vec{f}_{q-p} = (\hat{R}_p)\ \vec{\beta}\ ;(p = 1, 2, \ldots, q-1), \tag{19}$$

where

$$\hat{R}_p = \alpha^p - k_1\ \alpha^{p-1} - \ldots - Ik_p \tag{20}$$

This procedure requires that F have an inverse, that is, the determinant of F be nonzero. One can show by appropriate manipulation ofthe columns of the det F that

$$\det F = \det\ (\alpha^{q-1}\ \vec{\beta},\ \alpha^{q-2}\ \vec{\beta},\ \ldots, \alpha\vec{\beta},\ \vec{\beta}). \tag{21}$$

However the system described by equation (1) will be controllable only if the matrix with columns

$$\vec{\beta},\ \alpha\vec{\beta},\ .\ .\ .,\ \alpha^{q-1}\ \vec{\beta}\ \text{has rank } q.$$

The determinant of this matrix is identical to that in equation (21) so this procedure is possible only if the system is controllable.

In summary the following procedure is used in the design of a system for specified characteristic roots:

(1) Determine the characteristic polynomial of the original open loop system, eq. (5) and the coefficients $(k_1, k_2, \ldots k_q)$.

(2) Specify the closed loop characteristic roots, and compute the characteristic polynomial eq. (10) to obtain the coefficients $(h_1, h_2, \ldots h_q)$.

(3) Compute $\underline{n}$ from (11).

(4) Compute F and F^{-1} from (19).

(5) Solve for $\underline{m}$ from (17).

II DISCUSSION

If the plant is controllable the procedure in I provides a simple method for changing its characteristic polynomial. However for some physical systems there may be practical limitations. The feedback gains required by the vector $\underline{m}$ may be too large to be physically realized, may lead to a system in which saturation occurs or where the linear description eq. (1) is no longer adequate. Also the effects of variations in the system parameters have not been considered. That is, a small change in a parameter could lead to an unstable system. The effect of variations in the system parameters is discussed in the next section.

The presentation and description in I is in the time domain. Consider a discussion in the Laplace or "s" domain. Therefore take the Laplace transform of equation (1) and (2), assuming zero initial conditions, and solve for the transfer function relationship between the output and the input. It is

$$\frac{f_{1q}\ s^{q-1} + f_{1q-1}\ s^{q-2} + \ .\ .\ .\ f_{12}s + f_{11}}{s^q - k_1\ s^{q-1} - \ .\ .\ .\ - k_q} \qquad (22)$$

After feedback the system transfer function is

$$\frac{(f_{1q}\, s^{q-1} + \ldots f_{12}\, s + f_{11})\, c_1}{s^q - h_1\, s^{q-1} - \ldots - h_q} \tag{23}$$

From (23) one sees the basis for the choice of c_1. It determines the static gain. One also sees that the numerator polynomial is invariant under state variable feedback. Thus one can shift the poles of the system to any desired position without increasing the order of the system and without changing the zeros of the plant transfer function.

III EFFECTS OF CHANGES IN SYSTEM PARAMETERS

The purpose of this section is to present an expression which relates the effects of changes of the system parameters on the characteristic roots of the system. The expression appears in equation (40). In the development of the expression the main assumptions are that the characteristic roots are distinct and the changes are to a first approximation. The development begins with a discussion of certain properties of a square matrix. This discussion closely follows pages 82 to 89 in Gantmacher (4). The properties of interest are the characteristic polynomial of the matrix and the adjoint matrix. The matrix A is chosen in order that the expression and formulas developed can be readily related to Section I.

Consider a square q^{th} order matrix A with elements a_{ij}. The characteristic matrix of A is $I\lambda - A$. The determinant of the characteristic matrix is $g(\lambda)$ where

$$g(\lambda) = \det(I\lambda - A) = \lambda^q - h_1\lambda^{q-1} - h_2\lambda^{q-2} - \ldots - h_q. \tag{24}$$

The matrix $R(\lambda)$ is the adjoint matrix and

$$(I\lambda - A)\, R(\lambda) = g(\lambda)\, I. \tag{25}$$

A formula expressing the adjoint matrix $R(\lambda)$ in terms of the characteristic polynomial $g(\lambda)$ can be obtained. Using (24) form the difference $g(\lambda) - g(e)$ and note it is divisible by $\lambda - e$ without remainder. Thus

$$h(\lambda, e) = \frac{g(\lambda) - g(e)}{\lambda - e} = \lambda^{q-1} + (e - h_1)\lambda^{q-2} + (e^2 - h_1 e - h_2)\lambda^{q-3} + \ldots \tag{26}$$

is a polynomial in λ and e. The identity

$$(\lambda - e)\, h(\lambda, e) = g(\lambda) - g(e) \tag{27}$$

holds if λ is replaced by $I\lambda$ and e by A. Thus

$$(I\lambda - A)\ h\ (I\lambda, A) = g\ (I\lambda) \tag{28}$$

since by the Cayley - Hamilton Theorem $g(A) = 0$.

Comparing (28) and (25)

$$h\ (I\lambda,\ A) = R\ (\lambda). \tag{29}$$

From (26)

$$R\ (\lambda) = I\ \lambda^{q-1} + R_1\ \lambda^{q-2} + R_2\ \lambda^{q-3} + \ .\ .\ .\ + R_{q-1}, \tag{30}$$

where

$$R_p = A^p - h_1\ A^{p-1} - \ .\ .\ .\ .\ - I\ h_p \tag{31}$$

To the matrix A belongs the distinct characteristic roots λ_i (i = 1, 2, . . .q), the characteristic vectors $\vec{p}_i$, and the characteristic rows $\underline{t_i}$. Therefore

$$\Lambda = PAT = \begin{bmatrix} \lambda_1 & 0 & . & . & . & 0 \\ 0 & \lambda_2 & . & . & . & 0 \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ 0 & 0 & . & . & . & \lambda_q \end{bmatrix} . \tag{32}$$

One can also write

$$\lambda_i\ \vec{p}_i = A\ \vec{p}_i\ , \tag{33}$$

$$\lambda_i\ \underline{t_i} = \underline{t_i} A. \tag{34}$$

The characteristic roots λ_i will be changed by a small amount $\partial\lambda_i$ when there is a change ∂A in the matrix A. An expression for the change $\partial\lambda_i$ in terms of ∂A was first obtained by Jacobi. Bodewig (5) reports Jacobi's derivation. Jacobi's expression is

$$\partial\lambda_i = \underline{t_i}\ (\partial A)\vec{p}_i \tag{35}$$

Equation 35 expresses in a first approximation the correction of λ_i by the value of a bilinear form. However it is necessary to know both $\underline{t_i}$ and $\vec{p}_i$ and so in its present form it is practically useless for design purposes.

Equation (35) can be rewritten if one first lets

$$\partial A = (\partial \vec{a}_1, \partial \vec{a}_2, \dots, \partial \vec{a}_q). \tag{36}$$

Thus equation (35) can be written as

$$\partial \lambda_i = p_{1i} \underline{t}_i \partial \vec{a}_1 + p_{2i} \underline{t}_i \partial \vec{a}_2 + \dots \; p_{qi} \underline{t}_i \partial \vec{a}_q \, . \tag{37}$$

Now define the inner product of the matrices A and B as

$$A * B = \underline{a}_1 \vec{b}_1 + \underline{a}_2 \vec{b}_2 + \dots + \underline{a}_q \vec{b}_q \tag{38}$$

where $\underline{a}_i$ are the rows of A and $\vec{b}_i$ are columns of B.

Therefore

$$\partial \lambda_i = \vec{p}_i \underline{t}_i * \partial A. \tag{39}$$

An expression equivalent to equation (39) can be written in terms of the adjoint matrix. It is

$$\partial \lambda_i = \frac{h\,(I\lambda_i, A) * \partial A}{g'\,(\lambda_i)} \tag{40}$$

where $h\,(I\lambda_i, A)$ is the adjoint matrix evaluated at λ_i and $g'\,(\lambda_i)$ is the derivative of $g(\lambda)$ with respect to λ and evaluated at λ_i.

The expression in eq (40) is derived by starting with

$$\lim_{\lambda \to \lambda_i} (\lambda - \lambda_i)\,(I\lambda - A)^{-1} = \lim_{\lambda \to \lambda_i} (\lambda - \lambda_i)\,\frac{h\,(I\lambda\,, A)}{g\,(\lambda)}. \tag{41}$$

One then shows that

$$\lim_{\lambda \to \lambda_i} (\lambda - \lambda_i)\,(I\lambda - A)^{-1} = \lim_{\lambda \to \lambda_i} (\lambda - \lambda_i)\; P\,(I\lambda - \Lambda)^{-1} T = \vec{p}_i \underline{t}_i \tag{42}$$

and

$$\lim_{\lambda \to \lambda_i} (\lambda - \lambda_i)\,\frac{h\,(I\lambda, A)}{g\,(\lambda)} = \frac{h\,(I\lambda_i, A)}{g'\,(\lambda_i)}\,, \tag{43}$$

so that

$$\vec{p}_i \underline{t}_i = \frac{h\,(I\lambda_i\,, A)}{g'\,(\lambda_i)}\,. \tag{44}$$

An expression for $\partial \lambda_i$ can also be written in terms of the coefficients of the characteristic polynomial. This can be done by applying Newton's formula for the approximation to the roots of a polynomial to the characteristic polynomial $g\,(\lambda)$. If this is done the result is

$$\partial \lambda_i = \frac{\partial h_1 \lambda_i^{q-1} + \partial h_2 \lambda_i^{q-2} + \dots + \partial h_q}{g'\,(\lambda_i)} \tag{45}$$

where $\partial h_1, \ldots \partial h_q$ are changes in the coefficients $h_1, \ldots, h_q$. Formulas for ∂h_p $(p = 1, \ldots, q)$ are

$$\begin{aligned} \partial h_1 &= I * \partial A \\ \partial h_2 &= R_1 * \partial A \\ &\vdots \\ \partial h_p &= R_{p-1} * \partial A \end{aligned} \tag{46}$$

Equation (40) is not a satisfactory design expression to be used with Section I. One can be obtained by first noting that by using eq (13) in eq (8) and comparing with eq (15) that

$$A = F\,H\,F^{-1}. \tag{47}$$

Therefore an expression for the change in any root λ_i of the closed loop system of equation (15) to a change in the parameters of the system is

$$\partial \lambda_i = \frac{F\,h\,(I\lambda_i, H)\,F^{-1} * \partial A\,.}{g'(\lambda_i)} \tag{48}$$

Equation (48) is in a convenient form since it can be shown (see reference (6)) that the adjoint matrix $h\,(I\,\lambda_i, H)$ can be written as the outer product of two vectors. That is,

$$h\,(I\lambda_i, H) = \vec{r}_i\,\underline{s}_i \tag{49}$$

where the transpose of the vector $\vec{r}_i$ is

$$\vec{r}_i^{\,T} = (1, \lambda_i, \lambda_i^2, \ldots \lambda_i^{q-1}), \tag{50}$$

and

$$\underline{s}_i^{\,T} = \begin{bmatrix} s_1 \\ \cdot \\ \cdot \\ \cdot \\ s_{q-2} \\ s_{q-1} \\ s_q \end{bmatrix} = \begin{bmatrix} \lambda_i^{q-1} - h_1\lambda_i^{q-2} - \cdots - h_{q-1} \\ \cdot \\ \cdot \\ \cdot \\ \lambda_i^2 - h_1\lambda_i - h_2 \\ \lambda_i - h_1 \\ 1 \end{bmatrix} \tag{51}$$

(In reference (6) the procedure is to obtain h $(I\lambda_i, H)$ for a third order system and then to note that the result can be generalized. Equations (29) thru (31) and the Cayley - Hamilton Theorem are used to obtain $h(I\lambda_i, H)$ in the form of equations (50) and (51).)

It is possible to write s_i in terms of the characteristic roots. This can be done by using expressions for the coefficients of the characteristic polynomial in terms of the characteristic roots to show that

$$s_{q-1} = - \sum_{\substack{j=1 \\ j\neq i}}^{q} \lambda_i, \tag{52}$$

$$s_{q-2} = \sum_{\substack{j=1 \\ j\neq i}}^{q-1} \sum_{\substack{k=j+1 \\ k\neq i}}^{q} \lambda_j \lambda_k , \tag{53}$$

$$s_{q-3} = - \sum_{\substack{j=1 \\ j\neq i}}^{q-2} \sum_{\substack{k=j+1 \\ k\neq i}}^{q-1} \sum_{\substack{m=j+2 \\ m\neq i}}^{q} \lambda_j \lambda_k \lambda_m . \tag{54}$$

Equations (52) thru (54) can be extended in a similar manner so the following formulas can be written

s_q always equals 1,

s_{q-1} is the negative of the sum of the roots except for λ_i,

s_{q-2} is the product of the roots given in s_{q-1} taken two at a time,

s_{q-3} is the negative of the product of the roots given in s_{q-1} taken three at a time, and so on.

It should be noted that for a given λ_i one can write

$$g'(\lambda_i) = (-1)^{q+1} \left((\lambda_{i+1} - \lambda_i)(\lambda_{i+2} - \lambda_i)\ldots(\lambda_q - \lambda_i)(\lambda_1 - \lambda_i) (\lambda_2 - \lambda_i) \; (\lambda_3 - \lambda_i) \ldots \right). \tag{55}$$

As an illustrative example h $(I\lambda_1, H)$ and g' (λ_1) for a fourth order system are

$$\vec{r}_1^{\,T} = (1, \lambda_1, \lambda_1^2, \lambda_1^3) , \tag{56}$$

$$\underline{s}_1 = (-(\lambda_2 \lambda_3 \lambda_4), (\lambda_2 \lambda_3 + \lambda_2 \lambda_4 + \lambda_3 \lambda_4), - (\lambda_2 + \lambda_3 + \lambda_4), + 1), \tag{57}$$

and

$$g' (\lambda_1) = (-1)^3 (\lambda_2 - \lambda_1) (\lambda_3 - \lambda_1) (\lambda_4 - \lambda_1) \tag{58}$$

IV SENSITIVITY FUNCTIONS

Using equation (48) it is now possible to write expressions for sensitivity functions. The sensitivity function , S_b^a, is the ratio of the percentage change in "a" to the percentage change in "b" assuming all other changes are zero.

First however let ∂a_{kj} be the **kj**th element of ∂A, and since

$$a_{kj} = \alpha_{kj} + \beta_k m_j , \tag{59}$$

one has

$$\partial a_{kj} = \partial \alpha_{kj} + (\partial \beta_k) m_j . \tag{60}$$

Let F^{-1} be defined as the matrix G with columns $\vec{g}_1, \vec{g}_2, \ldots \vec{g}_q$ so that equation (17) can be used to obtain m_j where

$$m_j = \underline{n} \, \vec{g}_j . \tag{61}$$

Using equations (48), (60), and (61) one obtains

$$S_{\alpha_{kj}}^{\lambda_i} = \frac{\alpha_{kj} (\underline{f}_j \, \vec{r}_i)(\underline{s}_i \, \vec{g}_k)}{\lambda_i \, g'(\lambda_i)} \tag{62}$$

and after considerable algebraic manipulation

$$S_{\beta_k}^{\lambda_i} = \frac{\beta_k \, (\underline{s}_i \, \vec{g}_k)(\lambda_i^q - k_1 \lambda_i^{q-1} - \ldots - k_q)}{\lambda_i \, g'(\lambda_i)} \tag{63}$$

V ILLUSTRATIVE EXAMPLE

As an illustrative example take a plant represented by equations (1) and (2) with the matrix α and the vector $\vec{\beta}$ as shown in Table I. The plant transfer function is

$$\frac{s^2 + 6}{(s-1)(s+1)(s+2)} \tag{64}$$

TABLE I

ILLUSTRATIVE EXAMPLE DESIGN MATRICES

α	$\begin{matrix}0 & -1 & 0\\ -1 & -2 & -2\\ 1 & 0 & 0\end{matrix}$	$\vec{\beta}$	$\begin{matrix}1\\ 2\\ 3\end{matrix}$
K	$\begin{matrix}0 & 1 & 0\\ 0 & 0 & 1\\ 2 & 1 & -2\end{matrix}$	$\vec{r}_1^{\,T}$	$1 \quad \lambda_1 \quad \lambda_1^2$
		$\underline{s}_1$	$\lambda_2\lambda_3 \quad -(\lambda_2+\lambda_3) \quad 1$
		$g'(\lambda_1)$	$(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)$
F	$\begin{matrix}6 & 0 & 1\\ -2 & -7 & 2\\ -3 & 7 & 3\end{matrix}$	$35\,F^{-1}$	$\begin{matrix}5 & -1 & -1\\ 0 & -3 & 2\\ 5 & 6 & 6\end{matrix}$

Assume that β_2 and α_{12} vary. The design objective is to obtain a closed loop stable system which is comparatively insensitive to these changes.

Using the information in Table I the sensitivity functions are

$$S_{\beta_2}^{\lambda_1} = \frac{2\,(-\lambda_2\lambda_3 + 3\,(\lambda_2 + \lambda_3) + 6)\,(\lambda_1^3 + 2\lambda_1^2 - \lambda_1 - 2)}{35\lambda_1\, g'\,(\lambda_1)} \tag{65}$$

$$S_{\alpha_{12}}^{\lambda_1} = \frac{-1\,(2\,\lambda_1^2 - 7\,\lambda_1 - 2)\,(5\,\lambda_2\,\lambda_3 + 5)}{35\,\lambda_1\, g'\,(\lambda_1)} \tag{66}$$

Equations (65) and (66) are used to select the values of the closed loop characteristic roots. (It should be noted that the sensitivity functions for λ_2 and λ_3 have the same form as for λ_1. That is, $S_{\beta_2}^{\lambda_2}$ is obtained from equation (65) by replacing λ_1 by λ_2 and λ_2 by λ_1). The expression, in equation (65), $\lambda_1^3 + 2\,\lambda_1^2 - \lambda_1 - 2$ will be zero if $\lambda_1 = 1, -1$, or -2. Thus let $\lambda_2 = -1$ and $\lambda_3 = -2$ so that

$$S_{\beta_2}^{\lambda_2} = 0 \text{ and } S_{\beta_2}^{\lambda_3} = 0 \tag{67}$$

Now choose λ_1 so that $S_{\alpha_{12}}^{\lambda_1}$ is zero. Thus let $\lambda_1 = -0.2655$. With these values of the characteristic roots Table II results.

TABLE II

SENSITIVITY FUNCTIONS

i	1	2	3
λ_i	-0.265	-1	-2
$S^{\lambda i}_{\beta 2}$	1.365	0	0
$S^{\lambda i}_{\alpha 12}$	0	2.085	1.04

Thus for a 10% change in β_2 and a 10% change in α_{12} there is only a change of 13.65% in λ_1, a 20.85% change in λ_2 and a 10.4% change in λ_3.

Acknowledgement: The author wishes to acknowledge the most helpful discussion and comments of Professor Elmer G. Gilbert of the University of Michigan on the work reported in this paper.

REFERENCES

1. D. J. Povejsil and A. M. Fuchs, "A Method for the Preliminary Synthesis of a Complex Multi-Loop Control System," Trans. AIEE, Vol. 74, Part 2 pp. 129-34, July 1955.
2. R. W. Bass and P. Mendelson, "Aspects of General Control Theory," Final Technical Report AFOSR 2754 Contract AF 49(638)-1003 Aeronca Manufacturing Corp. Baltimore, Maryland, 1962.
3. R. E. Kalman, "When is a Linear Control System Optimal?", RIAS Technical Report 63-5, March 1963.
4. F. R. Gantmacher The Theory of Matrices, Vol. I, Chelsea Publishing Company, New York, N. Y., 1959.
5. E. Bodewig, Matrix Calculus, North Holland Publishing Company, Amsterdam, 1956.
6. B. S. Morgan, Jr., "The Synthesis of Linear Multivariable Systems By State Variable Feedback," PhD Thesis, University of Michigan, 1963.

THE SENSITIVITIES OF A NETWORK WITH MULTIPLE PARAMETER VARIATIONS AND THEIR MINIMIZATION*

Y. FU
Department of Electrical Engineering
University of Kansas
Lawrence, Kansas

and

S. S. YAU
Department of Electrical Engineering
Northwestern University
Evanston, Illinois

ABSTRACT

In this paper, the sensitivity of a network function and the pole-zero sensitivity with n-parameter variations are defined. These n parameters define an n-dimensional space, in which the sensitivities for a given set of the n-parameter deviations and the maximum sensitivity among all possible deviations of the n parameters are evaluated. The maximum sensitivities of different realizations of a network function are used for comparison as the qualities of the realizations. A technique for minimizing the sensitivity of a specific realization without changing its basic structure is given, and a correlation between the sensitivity of a network function and its pole-zero sensitivity is established.

I. Introduction

In network synthesis, we usually have many different realizations for a given network function F(s). The network function of a specific realization if a function of the network parameters, some or all of which may deviate from the designed values or nominal values in the physical network. Consequently, the network function of the physical network is different from the specified F(s). In order to study the effects of variations of a certain parameter x on the network function F(x, s) of a specific realization, Bode [1] first defined the sensitivity S_x^F of the network function F(x, s) as follows:

$$S_x^F = \left[\frac{d(\ln F)}{d(\ln x)}\right]^{-1}. \tag{1}$$

*This work was jointly supported by the Office of Scientific Research, U.S. Air Force and the U.S. Army Research Office, Durham, North Carolina. This paper was completed when Dr. Fu worked with Dr. Yau at the Northwestern University during the summer of 1963.

Mason [2] took the reciprocal of (1) as the definition of sensitivity in the study of feedback theory. In recent years, the utilization of the pole-zero approach in network theory has greatly been increased, and the pole-zero configuration plays a very important role in the network responses. In order to investigate the effects of the variations of certain parameter x on the locations of the poles and zeros of a network function, Truxal and Horowitz [3] defined the pole-zero sensitivity $S_x^{s_j}$ as follows:

$$S_x^{s_j} = \frac{ds_j}{dx} x, \tag{2}$$

where s_j is a pole or a zero of the network function F(x, s). Since then, a number of authors [4-10] have worked on the sensitivity problem in both network analysis and synthesis. However, a more important problem is to consider the sensitivity when several or all of the network parameters change simultaneously. Hakimi and Cruz [11] have suggested some sensitivity measures with multiple parameter variations, and K.S. Fu [12] has found the effects of variations of several parameters on the location of a pole or a zero. Goldstein and Kuo [13] extended Mason's definition to the multiparameter variation case. In this paper, we first define the sensitivity measures with n-parameter variations for both a network function and the pole-zero locations. These n parameters define an n-dimensional space, in which the sensitivities for a given set of the n-parameter deviations and the maximum sensitivity among all possible deviations of the n parameters are evaluated. The maximum sensitivities of different realizations of a network function are used for comparison as the qualities of the realizations. Because it is usually desired that a network is not sensitive with respect to its parameter deviations, a technique for minimizing the sensitivity for a given network realization is given. Finally, a correlation between the sensitivity of a network function and its pole-zero sensitivity is established.

II. The Definitions of Sensitivities

In order to show only the effects of variations of a parameter x upon a network function F(x, s), the sensitivity S_x^F may be defined as

$$S_x^F = \frac{dF}{dx} . \tag{3}$$

This definition can be made equivalent to that given by Mason by multiplying a factor of x/F, which is a known quantity.

In the case where the n parameters $x_1, x_2, \ldots, x_n$ vary simultaneously, the network function is explicitly expresses as $F(x_1, x_2, \ldots, x_n, s)$, and the n parameters $x_1, x_2, \ldots, x_n$ define an n-dimensional space. Let $\overline{x}$ be an ordered set of n real numbers $(x_1, x_2, \ldots, x_n)$, which is a point in the n-dimensional space, or can be considered as a vector from the origin to that point. Hence, we can write $F(x_1, x_2, \ldots, x_n, s)$ as $F(\overline{x}, s)$. Suppose the n parameters change from their nominal values by infinitesimal quantities $dx_1, dx_2, \ldots, dx_n$). These variations can be considered as a vector $\overline{dx} = (dx_1, dx_2, \ldots dx_n)$, whose magnitude is given by the equation,

$$\left|\overline{dx}\right| = \left[\sum_{i=1}^{n} (dx_i)^2\right]^{1/2}, \tag{4}$$

and the directional angles $\alpha_1, \alpha_2, \ldots, \alpha_n$ are given by the equations

$$\alpha_i = \cos^{-1} (dx_i/|\overline{dx}|), \; i = 1, 2, \ldots, n. \tag{5}$$

Then, the rate of change of the network function $F(\overline{x}, s)$ due to the deviation $\overline{dx}$ is its directional derivative in the direction of $\overline{dx}$, which is given by the expression

$$\frac{dF}{\overline{dx}} = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i} \cos \alpha_i . \tag{6}$$

For convenience, we difine the sensitivity $S_{\overline{dx}}^F$ of a network function $F(\overline{x}, s)$ with respect to the deviation $\overline{dx} = (dx_1, dx_2, \ldots, dx_n)$ as the rate of the change of $F(\overline{x}, x)$ due to $\overline{dx}$, that is

$$S_{\overline{dx}}^F = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i} \cos \alpha_i , \tag{7}$$

where α_i's are given in (5).

It is noted that the above definition of the sensitivity is not the ratio of the percentage change of $F(\overline{x}, s)$ and the percentage change of $\overline{x}$. We adopt this definition because the change of $F(\overline{x}, s)$ depends on the absolute change of each parameter x_i, not its percentage change. If we want to express the multiparameter sensitivity in terms of percentage changes, we only need to multiply a factor $\overline{x}/F$ or $(\sum_{i=1}^{n} x_i^2)^{1/2}/F$ to the above definition. If there is only one parameter varying, this product will be exactly the same as Mason's definition.

The zeros and poles of the network function $F(\overline{x}, s)$ are in general functions of the n parameters $x_1, x_2, \ldots, x_n$. Thus, the pole-zero sensitivity with respect to the deviation $\overline{dx}$ can be defined in the similar manner, that is

$$S^{s_j}_{\overline{dx}} = \sum_{i=1}^{n} \frac{\partial s_j}{\partial x_i} \cos \alpha_i , \tag{8}$$

where s_j is the zero or pole of $F(\overline{x}, s)$ we are concerned, and α_i's are given in (5).

III. The Maximum of the Sensitivity

It follows from the previous definitions that the sensitivity of a network function and the pole-zero sensitivity all depend on the direction of the deviation $\overline{dx}$. When we compare the qualities of different realizations of a specified network function $F(s)$, we do not care what direction of the deviation $\overline{dx}$ is, but we want to know the worst possible case. In other words, we are interested in the maximum rate of change of $F(\overline{x}, s)$ regardless how the parameters $x_1, x_2, \ldots, x_n$ deviate. Therefore, it is necessary to find the maximum of the sensitivity of $F(\overline{x}, s)$ among all possible directions of the deviation $\overline{dx}$. Since this maximum is independent of $\overline{dx}$, we denote it by $S^F_{\overline{x}}$.

Consider the gradient vector of $F(\overline{x}, s)$ in the n-dimensional space. It is given by the equation

$$\text{grad}\, F = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i} \hat{x}_i , \tag{9}$$

wnere $\hat{x}_i$ is the unit vector in the direction of x_i-axis. The unit vector in the direction of $\overline{dx}$ is

$$\frac{\overline{dx}}{|\overline{dx}|} = \sum_{i=1}^{n} \cos \alpha_i \hat{x}_i, \tag{10}$$

where α_i' s are given by (5). Comparing the dot product of grad F and $\overline{dx} / |\overline{dx}|$ with (7), we have

$$S^F_{\overline{dx}} = (\text{grad } F) \cdot \frac{\overline{dx}}{|\overline{dx}|} \tag{11}$$

It is clear that (11) reaches its maximum when $\overline{dx}$ is in the same direction of grad F. Thus, the maximum sensitivity $S^F_{\overline{x}}$ among all possible directions of $\overline{dx}$ is equal to the magnitude of the gradient vector of $F(\overline{x}, s)$. That is,

$$\begin{aligned} S^F_{\overline{x}} &= \text{grad } F \\ &= \left[\sum_{i=1}^{n} \left(\frac{\partial F}{\partial x_i} \right)^2 \right]^{1/2} \end{aligned} \tag{12}$$

It is noted that $S^F_{\overline{x}}$ given in (12) is a function of the complex frequency s. In order to find its value in the real frequency ω domain, we write $S^F_{\overline{x}}$ in the form

$$S^F_{\overline{x}} \ (s=j\omega) = \left[\sum_{i=1}^{n} \left| \frac{\partial F}{\partial x_i} \right|^2 \right]^{1/2} \tag{13}$$

As discussed previously, $S^F_{\overline{x}}$ is actually used for comparing the sensitivities of different realizations. This is demonstrated by the following examples. Suppose that the driving-point impedance Z(s) is specified as

$$Z(s) = \frac{s^2 + 3s + 1}{s(2s+1)} \tag{14}$$

It can be realized by Cauer' s method as well as Foster' s method. The two different realizations are shown in Figure 1.

For the Cauer' s realization, we have

$$Z(\overline{x}, s) = \frac{s^2 C_1 C_2 + s(g_2 C_1 + g_1 C_2 + g_2 G_2) + g_1 g_2}{s^2 C_1 C_2 (g_1 + g_2) + s C_1 g_1 g_2}, \tag{15}$$

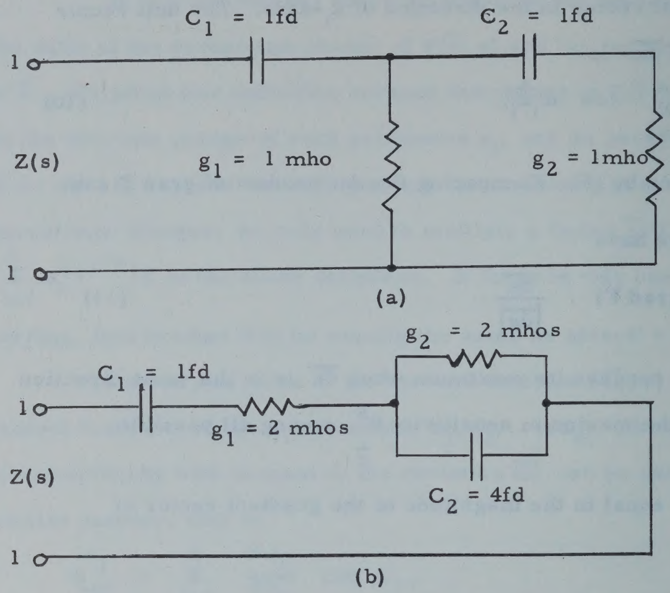


Figure 1. Two different realizations of the driving-point impedance $Z(s) = \frac{s^2+3s+1}{s(2s+1)}$. (a) Cauer's network, and (b) Foster's network.

where $\overline{x} = (C_1, C_2, g_1, g_2)$ whose nominal value is (1, 1, 1, 1,). Substituting (15) into (13) and letting $\overline{x}$ equal its nominal value, $S_{\overline{x}}^{Z}$ in the real frequency ω domain is given by the equation

$$S_{\overline{x}=(1,1,1,1)}^{Z}\Bigg|_{s=j\omega} = \left[\frac{\omega^6+20\omega^4+9\omega^2+1}{16\omega^6+16\omega^4+\omega^2}\right]^{1/2} \tag{16}$$

For the Foster's network,

$$Z(\overline{x}, s) = \frac{s^2C_1C_2+s(g_1C_1+g_2C_1+g_1C_2) + g_1g_2}{s^2C_1C_2g_1 + sC_1g_1g_2}\,, \tag{17}$$

where $\overline{x} = (C_1, C_2, g_1, g_2)$ whose nominal value is (1, 4, 2, 2). Similarly, we obtain

$$S_{\overline{x}=(1,4,2,2)}^{Z}\Bigg|_{s=j\omega} = \left[\frac{16\omega^6+265\omega^4+130\omega^2+16}{16(16\omega^6+8\omega^4+\omega^2)}\right]^{1/2} \tag{18}$$

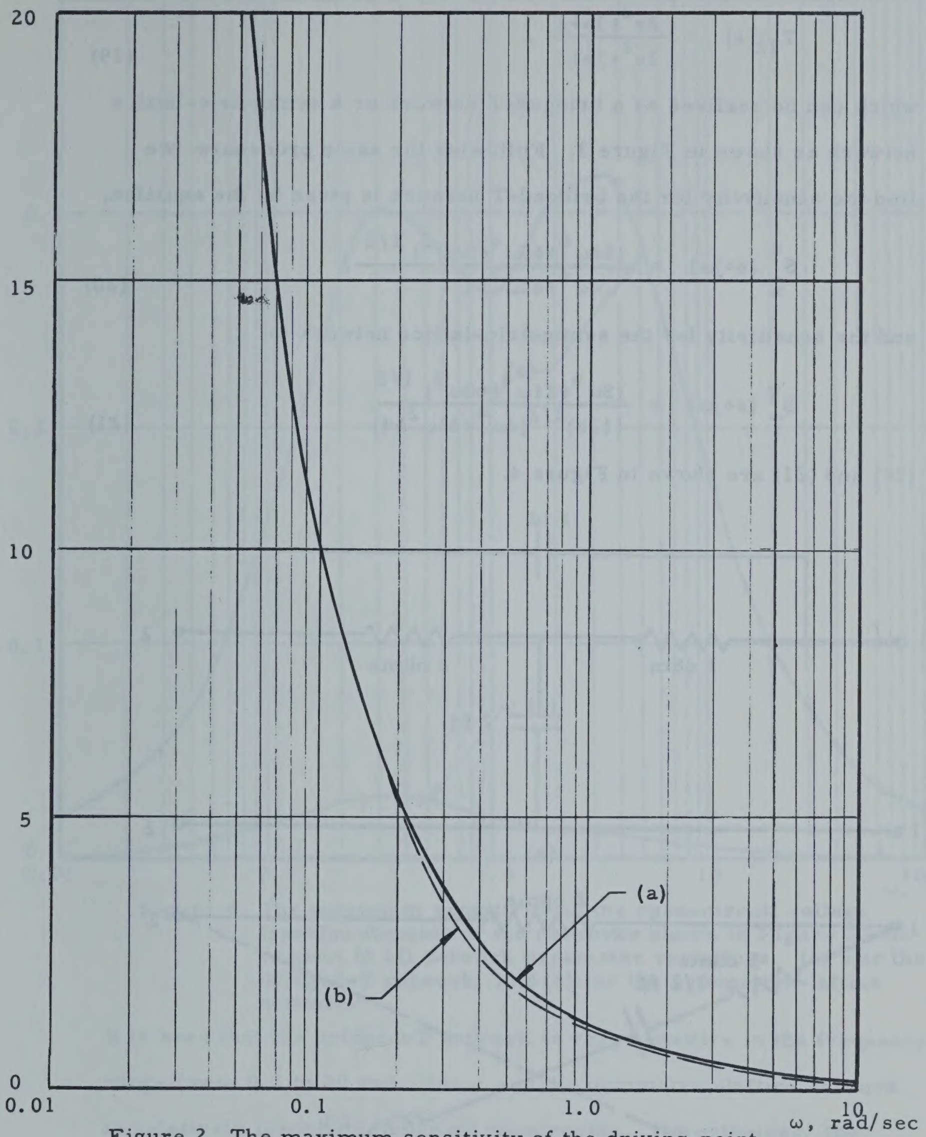


Figure 2. The maximum sensitivity of the driving-point impedance of the networks shown in Figure 1 with respect to all network parameter variations. (a) For the Cauer's network, and (b) for the Foster's network.

(16) and (18) are shown in Figure 2 for comparison. It can be seen that in this example Foster's realization is slightly less sensitive than Cauer's realization for all frequencies.

The next example illustrates that there are big differences between the sensitivities of various realizations of a given network function. Consider the open-circuit voltage transfer function

$$T_{12}(s) = \frac{2s^2+3s+2}{2s^2+7s+2}, \qquad (19)$$

which can be realized as a bridged-T network or a symmetric-lattice network as shown in Figure 3. Following the same procedure, we find the sensitivity for the bridged-T network is given by the equation,

$$S^{T}_{\underline{x}}\,(s=j\omega) = \frac{(64\omega^6+45\omega^4+56\omega^2)^{1/2}}{4\omega^4+41\omega^2+4}, \qquad (20)$$

and the sensitivity for the symmetric-lattice network is

$$S^{T}_{\underline{x}}\,(s=j\omega) = \frac{(8\omega^6+21\omega^4+68\omega^2)^{1/2}}{(128)^{1/2}(4\omega^4+41\omega^2+4)}, \qquad (21)$$

(20) and (21) are shown in Figure 4.

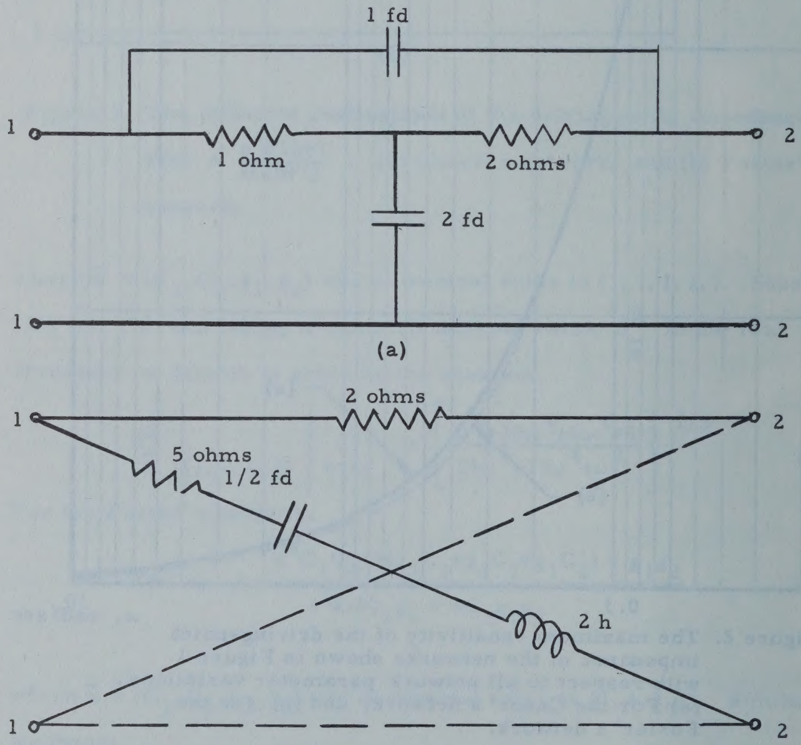


Figure 3. Two different realizations of the open-circuit voltage transfer function $T_{12}(s) = \frac{2s^2+3s+2}{2s^2+7s+2}$. (a) Bridged-T network, and (b) symmetric-lattice network.

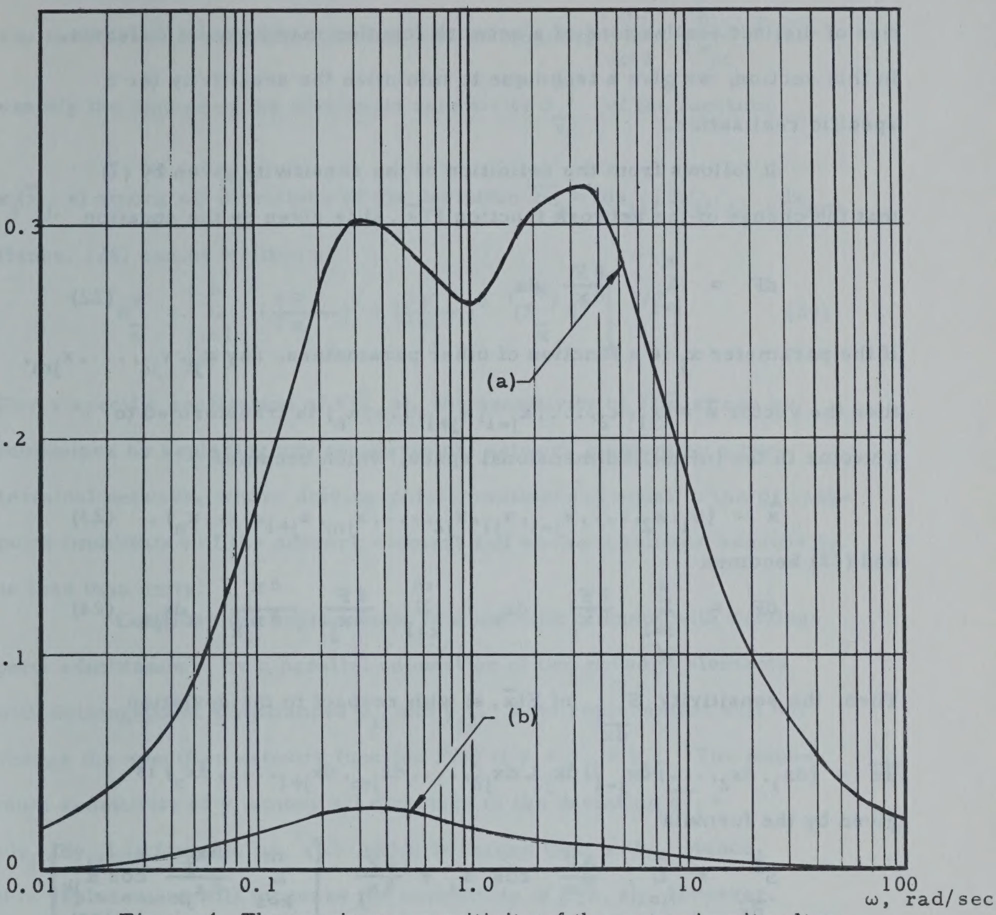


Figure 4. The maximum sensitivity of the open-circuit voltage transfer function of the networks shown in Figure 3 with respect to all network parameter variations. (a) For the bridged-T network, and (b) for the symmetric-lattice network.

It is seen that the bridged-T network is very sensitive in the frequency range from 0.1 to 10 rad./sec., and the symmetric-lattice network is relatively insensitive over all frequencies. Nevertheless, the symmetric-lattice network requires four resistors, two capacitors and two inductors; whereas the bridged-T network has only two resistors and two capacitors.

IV. Minimization of the Sensitivity

In most cases, it is necessary to design a network such that its network function is insensitive with respect to its parameter variations. We can see from the previous examples that the sensitivities of distinct realizations of a network function may be quite different. In this section, we give a technique to minimize the sensitivity for a specific realization.

It follows from the definition of the sensitivity given by (7) that the change of the network function $F(\overline{x}, s)$ is given by the equation

$$dF = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i} dx_i \tag{22}$$

If the parameter x_j is a function of other parameters, say $x_{j1}, x_{j2}, \ldots, x_{jm}$, then the vector $\overline{x} = (x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$ is transformed to a vector in the (n+m-1)-dimensional space, which becomes

$$\overline{x} = (x_1, x_2, \ldots, x_{j-1}, x_{j1}, x_{j2}, \ldots, x_{jm}, s_{j+1}, \ldots, x_n), \tag{23}$$

and (22) becomes

$$dF = \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{\partial F}{\partial x_i} dx_i + \sum_{k=1}^{m} \frac{\partial F}{\partial x_j} \frac{\partial x_j}{\partial x_{jk}} dx_{jk} \tag{24}$$

Then, the sensitivity $S^F_{\overline{dx}}$ of $F(\overline{x}, s)$ with respect to the deviation $\overline{dx} = (dx_1, dx_2, \ldots, dx_{j-1}, dx_{j1}, dx_{j2}, \ldots, dx_{jm}, dx_{j+1}, \ldots, dx_n)$ is given by the formula

$$S^F_{\overline{dx}} = \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{\partial F}{\partial x_i} \cos \alpha_i + \frac{\partial F}{\partial x_j} \left[\sum_{k=1}^{m} \frac{\partial x_j}{\partial x_{jk}} \cos \alpha_{jk} \right], \tag{25}$$

where

$$\begin{aligned} \alpha_i &= \cos^{-1} (dx_i / |\overline{dx}|), \\ \alpha_{jk} &= \cos^{-1} (dx_{jk} / |\overline{dx}|), \end{aligned} \tag{26}$$

$$\text{and } |\overline{dx}| = \left[\sum_{\substack{i=1 \\ i \neq j}}^{n} (dx_i)^2 + \sum_{k=1}^{m} (dx_{jk})^2 \right]^{1/2} \tag{27}$$

The maximum sensitivity $S^F_{\overline{x}}$ among all directions of $\overline{dx}$ becomes

$$S^F_{\overline{x}} = \left[\sum_{\substack{i=1 \\ i \neq j}}^{n} \left(\frac{\partial F}{\partial x_i} \right)^2 + \left(\frac{\partial F}{\partial x_j} \right)^2 \sum_{k=1}^{m} \left(\frac{\partial x_j}{\partial x_{jk}} \right)^2 \right]^{1/2} \tag{28}$$

We may interpret the parameter x_j as any two terminal subnetwork or any two-terminal network element, and the quantity $\sum_{k=1}^{m} (\frac{\partial x_j}{\partial x_{jk}})^2$ is exactly the square of the maximum sensitivity $S_{\overline{x}_j}^{x_j}$ of the function $x_j(\overline{x}_j, s)$ among all directions of the deviation $\overline{dx}_j = (dx_{j1}, dx_{j2}, \ldots, dx_{jm})$. Hence, (28) can be written as

$$S_{\overline{x}}^{F} = \left[\sum_{\substack{i=1 \\ i \neq j}}^{n} (\frac{\partial F}{\partial x_i})^2 + (\frac{\partial F}{\partial x_j})^2 \quad (S_{\overline{x}_j}^{x_j})^2 \right]^{1/2} \tag{29}$$

For a specific realization of $F(\overline{x}, s)$, the sensitivity of $F(\overline{x}, s)$ can be minimized by replacing any two-terminal network element by a two-terminal network, whose driving-point immittance is equal to the driving-point immittance of the network element and whose maximum sensitivity is less than unity.

Consider the replacement of a network element with driving-point admittance y_j by a parallel connection of two network elements with driving-point admittances y_{j1} and y_{j2}. This replacement will not change the specified network function F(s) if $y_j = y_{j1} + y_{j2}$. The maximum sensitivity of y_j among all direction of the deviation $\overline{dy}_j = (dy_{j1}, dy_{j2})$ is found to be $\sqrt{2}$, which is larger than unity. Hence, this replacement will increase the sensitivity of $F(\overline{x}, s)$. However, if we consider the parameters as impedances instead of admittances, then $y_j = 1/z_j = 1/z_{j1} + 1/z_{j2}$ and the maximum sensitivity of y_j among all directions of the deviation before and after this substitution are $\left|1/z_j\right|^2$ and

$$\left[\left| \frac{1}{z_{j1}} \right|^4 + \left| \frac{1}{z_{j2}} \right|^4 \right]^{1/2} \tag{30}$$

respectively. It is noted that (30) cannot be larger than $\left|1/z_j\right|^2$ because the relation $1/z_j = 1/z_{j1} + 1/z_{j2}$ must hold. Furthermore, (30) reaches its minimum, which is $1/2 \sqrt{2}$, when $z_{j1} = z_{j2} = 2z_j$. This implies that the sensitivity is reduced. We see that the sensitivity being increased or decreased for the parallel conncetion depends on whether the parameters are considered as admittances or impedances. It can easily be shown that the above situation is reversed for the substitution of a network element by its equivalent series connection. In order to take the advantage of both parallel connection and series connection and compensate their disadvantage, we replace a network element by four identical elements connected in the bridge form as shown in Figure 5.

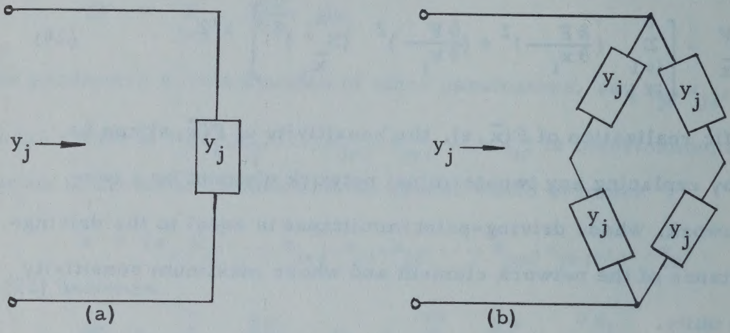


Figure 5. (a) A two-terminal network element or two-terminal subnetwork:. (b) Its equivalent balanced bridge network. The replacement of (a) by (b) reduces the sensitivity by one half.

In this case, we find that $S_{\overline{x}_j}^{x_j} = 1/2$ no matter whether the parameters are considered as driving-point admittances or driving-point impedances. If we replace every network element in a specific realization by such a bridge network, the sensitivity will be recduced to half of the original one. However, the number of elements in the modified network is four times as many as the original one. Consequently, the cost for this midification is also four times as the original cost. The modified network of the bridged-T network shown in Figure 3(a) is illustrated in Figure 6.

It follows from the same argument that the pole-zero sensitivity for this kind of modification is also reduced by half of the original one.

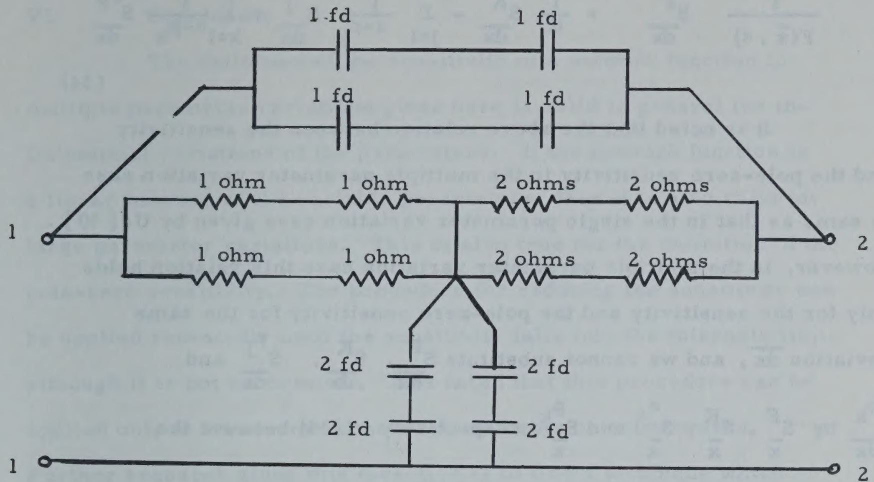


Figure 6. The modification of the bridged-T network shown in Figure 3(a). The sensitivity of this network is only one half of the original one.

V. The Correlation Between Sensitivity and The Pole-Zero Sensitivity

In order to find the correlation between the sensitivity of a network function $F(\overline{x}, s)$ and the pole-zero sensitivity, $F(\overline{x}, s)$ is written in the form,

$$F(\overline{x}, s) = K \frac{\prod_{j=1}^{\alpha} (s-z_j)}{\prod_{k=1}^{\beta} (s-p_k)}, \tag{31}$$

where K, z_j' s and p_k' s are functions of $\overline{x}$. Taking the logarithmic derivative of both sides, we have

$$\frac{dF(\overline{x}, s)}{F(\overline{x}, s)} = \frac{1}{K} \sum_{i=1}^{n} \frac{\partial K}{\partial x_i} dx_i - \sum_{j=1}^{\alpha} \frac{1}{s-z_j} \left(\sum_{i=1}^{n} \frac{\partial z_j}{\partial x_i} dx_i \right) + \sum_{k=1}^{\beta} \frac{1}{s-p_k} \left(\sum_{i=1}^{n} \frac{\partial p_k}{\partial x_i} dx_i \right) \tag{32}$$

Since

$$dF(\overline{x}, s) = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i} dx_i , \tag{33}$$

dividing both (32) and (33) by $|\overline{dx}|$, and substituting the definitions of the sensitivity and the pole-zero sensitivity given in (17) and (18) yields

$$\frac{1}{F(\bar{x}, s)} S^{F}_{\overline{dx}} = \frac{1}{K} S^{K}_{\overline{dx}} - \sum_{j=1}^{\alpha} \frac{1}{s-z_j} S^{z_j}_{\overline{dx}} + \sum_{k=1}^{\beta} \frac{1}{s-p_k} S^{p_k}_{\overline{dx}} \tag{34}$$

It is noted that the above relation between the sensitivity and the pole-zero sensitivity in the multiple parameter variation case is same as that in the single parameter variation case given by Ur [10]. However, in the multiple parameter variation case this relation holds only for the sensitivity and the pole-zero sensitivity for the same deviation $\overline{dx}$, and we cannot substitute $S^{F}_{\overline{dx}}$, $S^{K}_{\overline{dx}}$, $S^{z_j}_{\overline{dx}}$ and $S^{p_k}_{\overline{dx}}$ by $S^{F}_{\bar{x}}$, $S^{K}_{\bar{x}}$, $S^{z_j}_{\bar{x}}$ and $S^{p_k}_{\bar{x}}$ respectively in (34) because the gradients of $F(\bar{x}, s)$, $K(\bar{x})$, $z_j(\bar{x})$ and $p_k(\bar{x})$ in general have different directions. Therefore, only the components of $S^{F}_{\bar{x}}$ can be expressed in terms the corresponding components of the variations of K, z_j' s and p_k' s. This can be done by collecting the coefficients of dx_i in (32). That is,

$$\frac{1}{F(\bar{x}, s)} \frac{\partial F}{\partial x_i} = \frac{1}{K} \frac{\partial K}{\partial x_i} - \sum_{j=1}^{\alpha} \frac{1}{s-z_j} \frac{\partial z_j}{\partial x_i} + \sum_{k=1}^{\beta} \frac{1}{s-p_k} \frac{\partial p_k}{\partial x_i} \tag{35}$$

for i=1, 2, . . . , n. Substituting (35) into (12), we have

$$\frac{1}{F(\bar{x}, s)} S^{F}_{\bar{x}} = \left[\sum_{i=1}^{n} \left(\frac{1}{K} \frac{\partial K}{\partial x_i} - \sum_{j=1}^{\alpha} \frac{1}{s-z_j} \frac{\partial z_j}{\partial x_i} + \sum_{k=1}^{\beta} \frac{1}{s-p_k} \frac{\partial p_k}{\partial x_i} \right)^2 \right]^{1/2} \tag{36}$$

It is clear that $\frac{\partial K}{\partial x_i}$, $\frac{\partial z_j}{\partial x_i}$, and $\frac{\partial p_k}{\partial x_i}$ are the rates of variations of $K(\bar{x})$, $z_j(\bar{x})$, and $p_k(\bar{x})$ in the direction of x_i-axis. Hence, (36) shows the relation between $S^{F}_{\bar{x}}$ and the rates of variations of the pole-zero locations and the constant multiplier $K(\bar{x})$ in n components.

It follows from (34) and (36) that the variations of the poles and zeros close to the ω-axis dominate the sensitivity $S^{F}_{\overline{dx}}$ and its maximum $S^{F}_{\bar{x}}$. This situation is very important in the study of linear control systems because the poles and zeros close to the w-axis govern the performance of the system. They are usually called the dominant poles (zeros) [14].

VI. Conclusion

The definition of the sensitivity of a network function for multiple parameter variations given here is valid in general for infinitesimal variations of the parameters. If the network function is a linear function of the variable parameters, then it is also valid for large parameter variations. This is also true for the definition of the pole-zero sensitivity. The procedure for reducing the sensitivity can be applied repeatedly until the sensitivity falls into the tolerable limit, although it is not economical. It is noted that this procedure can be applied only to two-terminal netwrok elements or subnetworks. Further research along this direction is to find a technique which is applicable to three-or four-terminal network elements or subnetworks. This procedure does not change the basic structure of the original network. Since the sensitivity of distinct realizations may be quite different as shown in the previous examples, we can exhaust all possible synthesis techniques to obtain all different realizations and choose the least sensitive one. However, whether there is an economical and systematic way to introduce some redundant elements for reducing the sensitivity needs further exploration.

ACKNOWLEDGMENT

The authors wish to express their thanks to Dr. S. L. Hakimi for helpful discussions and Mr. K. C. Wang for programming the numerical examples.

REFERENCES

1. H. W. Bode, Network Analysis and Feedback Amplifier Design, D. Van Nostrand Co., Inc., Princeton, New Jersey, Chapter 4; 1945.
2. S. J. Mason, "Feedback Theory - some properties of signal flow graphs," Proc. IRE, Vol. 41, pp. 1144-1156; Sept., 1953.
3. J. G. Truxal and I. M. Horowitz, "Sensitivity considerations in active network synthesis," Proc. Second Midwest Symp. on Circuit Theory, Michigan State University, East Lansing, Michigan, pp. 6-1 to 6-11; December, 1956.
4. F. F. Kuo, "Pole-zero sensitivity in network functions," IRE Trans. on Circuit Theory, Vol. CT-5, No. 4, p. 372; December, 1958.
5. F. F. Kuo, "Sensitivity of transmission zeros in RC network synthesis," IRE Intern'l Convention Record, New York, N. Y., Part 2, pp. 18-22; March, 1959.
6. J. J. Mikulski, "A correlation between classical and pole-zero sensitivity," Tech. Note No. 5, Contract No. AF49(638)-63, Electrical Engineering Research Laboratory, University of Illinois; 1959.
7. I. M. Horowitz, "Optimization of negative-impedance conversion methods of active RC synthesis," IRE Trans. on Circuit Theory, Vol. CT-6, No. 3, pp. 296-303; September, 1959.
8. D. A. Calahan, "Notes on the Horowitz optimization procedure," IRE Trans. on Circuit Theory, Vol. CT-6, pp. 352-354; Sept., 1960.
9. D. A. Calahan, "Sensitivity minimization in active RC synthesis," IRE Trans. on Circuit Theory, Vol. CT-9, No. 1, pp. 38-42; March, 1961.
10. H. Ur, "Root locus properties and sensitivity relations in control systems," IRE Trans. on Automatic Control, Vol. AC-5, No. 1, pp. 57-65; January, 1960.
11. S. L. Hakimi and J. B. Cruz, Jr., "Measures of sensitivity for linear systems with large multiple parameter variations," IRE WESCON Convention Record, Part 2, pp. 109-115; Aug., 1960.
12. K. S. Fu, "Sensitivity of a linear system with variations of one or several parameters," IRE Trans. on Circuit Theory, Vol. CT-7, No. 3, pp. 348-349; September, 1960.
13. A. J. Goldstein and F. F. Kuo, "Multiparameter sensitivity," IRE Trans. on Circuit Theory, Vol. CT-8, No. 2, pp. 177-178; June, 1961.
14. J. G. Truxal, Automatic Feedback Control System Synthesis, McGraw-Hill Book Co., Inc., New York, N. Y., Chapter 4; 1955.

DEGENERACY IN N-PORT NETWORKS

G. BIORCI, P. P. CIVALLERI and G. FIORIO
Istituto Elettrotecnico Nazionale Galileo Ferraris
Torino, Italy

Abstract

The topological structure of an n-port network may be such that, independently on the values of the branch conductances, its port conductances are not all independent, but there exist one or more equations of functional dependance between them: such a network is called degenerate. After a survey on basic properties of general degeneracies, one simple but important class of them is completely investigated: those in which equations of functional dependance are linear; rules are given for deducing these equations by simple observation of the topological structure of the network. Non-linear degeneracies are also examined in the case of n-ports based on $(n + 2)$ - node networks.

1.- Introduction

The problem of realizing an n-th order real symmetrical matrix G as the short-circuit conductance matrix of a resistive n-port based on a network with more than $(n + 1)$ nodes is still unsolved. After a suggestion of P. Slepian (1), who suspected that the analysis of n-ports is not so well known as it should in order to attack the problem of the synthesis, we have started a wide investigation on those properties of single-element-kind networks which appeared of high relevance to that aim.

The synthesis of a given G matrix involves conceptually three steps: the choice of a network structure, the choice of n ports on it, the computation of the branch values in such a way that the self and mutual conductances of those ports have the same values as the corresponding elements of the matrix. A network structure on which n ports have been chosen will be called an n-port structure.

An n-port structure will be said degenerate if, and only if, there exist one or more functions of some (or possibly all) mutual and possibly self conductances among the ports, which are identically equal to zero for all possible values of the branch conductances; in other words, an n-port structure is called degenerate when and only when its conductances are functionally dependent.

Now, at least at the moment, it does not seem a worth investigation to think of realization techniques which lead to the result only because the matrix elements take particular numerical values, in such a way as to fulfil relations of the previously said type. But this amounts to recognize that degenerate n-port structures have to be disregarded when one is concerned with the choice of a structure on which a given matrix should be realized.

Obviously, degeneracy could be investigated for any particular n-port structure, by writing down the equations which express the port conductances as a function of the branch conductances, and then by computing the rank of the Jacobian matrix of such equations; but this method could hardly give some general information. Furthermore, degeneracy is a property of the n-port structure, since it is independent of the values of the branch conductances, hence it is a topological property which can be investigated in a purely geometrical way.

The aim of this work is to present the achievements we have got in the field of the degeneracy.

2.- Definitions and basic properties

In the following, we shall indicate by n the number of ports of the network. We assume that the graph of the ports contains no loops: in fact, the port voltages must be independent. However, in general, it is not connected, but is composed of (p + 1) separate parts, each of which is a tree.

We shall limit our considerations to networks in which each node belongs to one port at least: in fact, if there were a node not belonging to any port, it could be eliminated by a star-mesh transformation without affecting the performance of the network at the ports. Therefore the number of nodes is $\underline{n} + \underline{p} + 1$.

We also neglect disconnected, or hinged, networks, since each separate, or hinged part of the latter can be studied separately from the other ones. It follows from this that each separate part of the graph of the ports is connected to at least another one. Therefore it is possible to add to the given $\underline{n}$ ports other $\underline{p}$ ports, which we call additional, numbered from $(\underline{n} + 1)$ to $(\underline{n} + \underline{p})$, choosing each of them across two nodes actually connected by a branch (fig. 1).

Hence every $\underline{n}$-port can be thought of as an $(\underline{n} + \underline{p})$-port in which the $\underline{p}$ additional ports are kept permanently open-circuited. Since this point of view will be very useful in the following, we may define it more precisely. We use the symbol $G_{ij}^{(r)}$ to indicate the short-circuit mutual conductance between the ports $\underline{i}$ and $\underline{j}$, with the additional ports $\underline{n} + 1$, $\underline{n} + 2$, ... $\underline{n} + \underline{r}$ open-circuited and the additional ports $\underline{n} + \underline{r} + 1$, $\underline{n} + \underline{r} + 2$, ... $\underline{n} + \underline{p}$ short-circuited.

The following obvious relationship holds:

$$G_{ij}^{(r+1)} = G_{ij}^{(r)} - \frac{G_{i,\, n+r+1}^{(r)}\, G_{j,\, n+r+1}^{(r)}}{G_{n+r+1,\, n+r+1}^{(r)}} \qquad (1)$$

which comes from the elimination of the voltage V_{n+r+1} in the equations of the $\underline{n}$-port, after the condition $I_{n+r+1} = 0$ has been imposed. Now, we can prove the following

Theorem 1: a necessary condition in order that the $G_{ij}^{(r+1)}$'s are subjected to one or more equations of functional dependence, is that the corresponding $G_{ij}^{(r)}$'s are subjected to the same equations.

The proof is very simple: since the previously said equations hold identically for any values of the branch conductances, let the conductance of the branch in parallel to the port n+r+1 be made to tend towards infinity: since such a conductance appears only in $G^{(r)}_{n+r+1,\ n+r+1}$, the fractional terms in eqs. (1) converge to zero, and the $G^{(r+1)}_{ij}$'s approach the $G^{(r)}_{ij}$'s: hence any functional relation between the former goes into the same relation between the latter.

The theorem can be stated in a more intuitive way: if an n-port is degenerate, any (n + 1)-port obtained from it including an additional port among the working ones is also degenerate.

By inverting the previous proposition, one obtains the

Theorem 2: a sufficient condition in order that the $G^{(r+1)}_{ij}$'s are functionally independent is that the $G^{(r)}_{ij}$'s are also independent.

That is, if the port conductances of an n-port are not subjected to any functional relationship, the same holds for the port conductances of any (n - 1)-port obtained from the latter by leaving a port permanently open-circuited.

An interesting consequence of theorem 2 is expressed from the following

Corollary: an n-port structure based on a full network, that is on a network in which there is a branch between any pair of nodes, is never degenerate.

In fact, let us add p additional ports to the said structure and include all of them among the working ones: we obtain an (n + p)-port based on a full network with n + p + 1 nodes, the port conductances of which will be consequently indicated by $G^{(o)}_{ij}$. We can use the results on the networks of this type recently achieved by several Authors (2) (3) (4): the $G^{(o)}_{ij}$'s are linear functions of the branch conductances, and the Jacobian matrix (which is square, in this particular case) has rank equal to its order. Hence the $G^{(o)}_{ij}$'s are independent and, after Theorem 2, the $G^{(p)}_{ij}$'s are also.

Another remarkable but obvious consequence of the above properties is that degeneracy can be produced only by lack of branches in the network.

Now we prove a theorem which has not a particular significance for the following treatment, but may be useful to simplify the search for degeneracies in an n-port structure.

Theorem 3: if an n-port structure is degenerate, all n-port structures obtained from it by arbitrarily rearranging the ports within each separate part of their graph are also degenerate. The converse is also true, in the sense that a non--degenerate n-port structure remains non-degenerate after such a transformation.

Let the equations of the n-port be written down as:

$$I_i = \sum_{j=1}^{n} G_{ij}^{(p)} V_j \qquad (i = 1, 2, \ldots, n) \qquad (2)$$

where the I_i's and the V_j's are respectively the port currents and voltages. Let now the ports be rearranged arbitrarily within each separate part of their graph: this amounts evidently to introduce new port voltages $V_{j'}$, and currents $I_{i'}$. Each one of the old voltages V_j is linearly expressible in terms of only those $V_{j'}$'s which are pertinent to ports in the same separate part: that is the equations hold:

$$V_j = \sum_{j'=1}^{n'} A_{jj'} V_{j'} \qquad (j = 1, 2, \ldots, n) \qquad (3)$$

where coefficients $A_{jj'}$ are 0 if j and j' are indifferent separate parts, are ± 1 or 0 if they are in the same part according well known rules.

On the other hand, the law of conservation of energy imposes that the new currents $I_{i'}$ are related to the old ones I_i by the transposed transformation of the (3):

$$I_{i'} = \sum_{i=1}^{n} A_{i'i} I_i \qquad (i' = 1', 2', \ldots, n') \qquad (4)$$

By substituting eqs. (3) and (4) into eqs. (2), and equating the coefficient of $V_{j'}$ in the i'-th equation to $G^{(p)}_{i'j'}$ one finds:

$$G^{(p)}_{i'j'} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i'i} G^{(p)}_{ij} A_{jj'} \qquad (5)$$

that is the well known transformation formula valid without limitations for an n-port based on (n + 1) nodes.

Now the linear equations (5) constitute obviously a nonsingular transformation of the $G^{(p)}_{ij}$'s into the $G^{(p)}_{i'j'}$'s: in fact all the previous talk could be repeated by interchanging the new and the old ports and hence getting to invert eqs. (5); but such a transformation neither introduces new functional relationships between the port conductances nor can possibly cancel the existing ones. Hence the theorem is proved.

Theorem 3 shows that being an n-port structure degenerate or non-degenerate is independent on the port structure within each separate part of their graph. One could profit of such a property in the examination of an n-port from the point of view of degeneracy, by changing the port structure in each separate part into a simpler one, for example into a starlike tree.

3.- Degeneracy in n-port structures with p = 0

When the graph of the ports is connected, the problem of degeneracy is particularly simple, since the only possible non-degenerate n-port structures are those based on the full network. In fact, the number of nodes being (n + 1) the number of branches of the latter is $n(n+1)/2$, exactly equal to the number of port conductances. Therefore, if the network does not include all possible branches, every n-port structure based on it is degenerate. In order to find the equation of degeneracy resulting from one missing branch, one could solve the set of equations which express the port conductances in terms of the branch ones for the full network:

$$G_{ij}^{(o)} = \sum_h A_{ih} A_{jh} g_h . \tag{6}$$

This is always possible since eqs. (6) are linear and independent. One obtains

$$g_h = \sum_{ij} a_{hij} G_{ij}^{(o)} \tag{7}$$

where the coefficients a_{hij} hold ± 1 or 0 (5). If the missing branch is r, the equation of degeneracy is obviously:

$$\sum_{ij} a_{rij} G_{ij}^{(o)} = 0 \tag{8}$$

This proves that if the graph of the ports is connected, all equations of degeneracy are linear.

Since any functional relation between the port conductances $G_{ij}^{(p)}$ of an n-port structure based on $n + p + 1$ nodes implies, by Theorem 1, an analogous relation between the port conductances $G_{ij}^{(o)}$ of the $(n + p)$-port structure obtained by adding the p additional ports, it is clear that the latter must split into two or more linear equations: this is not true, however, at least in general, for the former.

A deeper insight in the form of eqs. (8) can be very useful for the following treatment. In a recent work (6), one of us has shown that eqs. (8) can be obtained directly from the tree structure, that is without going through the solution of eqs. (6). Therefore, one can easily use the result of that work to establish the rule for writing down the equation of degeneracy resulting from one missing branch. The rule sounds: if the missing branch is connected between the nodes a and b, the equation of degeneracy is obtained by summing up all those port conductances G_{ij}, which have one index referred to a port incident to node a, and the other one referred to a port incident to node b; the sign of the element G_{ij} is to be taken: positive, if both ports i and j are oriented toward the corresponding nodes a and b or both oriented in the opposite direction; negative, if one of the two ports points toward its corresponding node and the other away.

Using this rule, one can easily write down the equations of degeneracy produced by the missing branches. A very interesting property is the following: the matrix elements G_{ij} contained in an equation of degeneracy cover a rectangular portion of the G matrix. In other words, the degeneracy of this type can be thought of in this way: if we feed with voltage generators having e.m.f.'s equal to ± 1 V all ports incident to node a, the sum of the currents in the ports connected to node b is equal to zero.

Fig. 2 shows an example.

4.- Linear degeneracies

Let us consider an n-port structure in which the graph of the ports is made of p + 1 separate parts, with p different from zero.

A limited but important class of degeneracies is that of the linear ones: that is the equations of functional dependance between elements of the conductance matrix are $\sum G_{ij} = 0$.

We prove firstly the following

Theorem 4: a sufficient condition in order that an n-port structure be linearly degenerate is that, when feeding some ports $j_1\ j_2\ j_3\ \dots\ j_p$ with voltage generators of ± 1 V, there exists a closed surface cutting only some ports $i_1\ i_2\ i_3\ \dots$

... i_q and branches whose current sum is identically (i.e. for all positive values of the conductances of all branches of the network) equal to zero.

In fact, if the currents entering into the surface are counted for example as positive, while the coming out ones are assumed as negative, Kirchhoff's first law yields the equation:

$$\sum_{i=i_1}^{i_q} I_i = 0 \qquad (9)$$

where the sum covers only the currents in ports i_1 i_2 .... i_q since the sum of the currents in the other branches cutted by the surface is identically equal to zero. In (9) each term has positive sign if the corresponding port points into the surface, negative in the opposite case. But

$$I_i = \sum_{j=j_1}^{j_p} G_{ij}^{(p)} \qquad (10)$$

where each term of the sum is affected by the sign plus or minus, according the corresponding port $\underline{j}$ is fed by a voltage generator of + 1 V or - 1 V.

By combining equations (9) and (10) one obtains:

$$\sum_{i=i_1}^{i_q} \sum_{j=j_1}^{j_p} G_{ij}^{(p)} = 0 \qquad (11)$$

which is a linear equation between the $G_{ij}^{(p)}$'s. An equation such as the (11) will be said rectangular. In fact, if one orders the rows and the columns of the G matrix in such a way as to have the rows corresponding to ports $\underline{i}$ in the first $\underline{q}$ places and the columns corresponding to ports $\underline{j}$ in the first $\underline{p}$ places, eq. (11) is obtained by summing all the matrix elements in the so obtained ($\underline{p}$ x $\underline{q}$)-rectangle: the signs of the $G_{ij}^{(p)}$'s are established according the following rule: let eachone of the p columns be affected by the sign plus or minus, according the corresponding port is fed by a voltage generator of + 1 V or − 1 V; let eachone of the q rows be affected by the sign plus or minus, according the corresponding port is entering or coming out from the surface: then the sign of each $G_{ij}^{(p)}$ is obtained as the product of the sign of row i and column j.

It is to be noted that some ports i may coincide with some ports j.

In the following, an equation as the (11) will be termed rectangular if and only if it fulfils also the above sign rule. The degeneracy involved by rectangular equations is termed rectangular degeneracy.

In order to understand the mechanism by which the conditions of Theorem 4 may be realized, it is convenient to examine firstly some simple cases.

Let an n-port be considered and let all its ports be short-circuited: a network is obtained, in which each node corresponds to a separate part of the graph of the ports. If such a network is disconnected or hinged, the n-port structure is affected by a rectangular degeneracy. In fact, if one feeds a port j by a voltage generator (fig. 3), the current in every port i separated from j by the hinge is identically zero: hence we have the simplest rectangular equation $G_{ij} = 0$.

Another simple case is that of two or more ports in series (fig. 4). In fact, when one feeds a port j the currents in the series-connected ports i and k are identically equal and hence the rectangular equation yields $G_{ij} - G_{kj} = 0$.

Hence we shall limit our considerations, in the following, to n-port structures not belonging to one or both the above classes.

Suppose a choice of ports j has been found, such that, when feeding in the said way these ports, there exists a surface cutting some ports i and branches whose currents have sum equal to zero. One may ask under what conditions this is due to the fact that the current vanishes identically in eachone of these branches. If a branch is internal, i.e. if its terminal nodes belong to the same separate part of the graph of the ports, the current flowing in it, is in general:

$$I = g \sum V_r \qquad (V_r = \pm 1 \text{ V}) \qquad (12)$$

where g is the conductance of that branch and $\sum V_r$ is the algebraic sum of the voltages of the generators inserted on the path of ports connecting its terminals. Then I may be identically equal to zero if and only if $\sum V_r = 0$, that is if the ports of that path are all short-circuited or an even number of them is fed, one half by generators pointing towards one terminal node, the other half by generators pointing towards the other.

If one considers, on the other hand, a connecting branch b, that is a branch the terminal nodes of which are in different separate parts of the graph of the ports, one easily sees that, under the assumed hypotheses, the current in it cannot identically vanish.

In fact suppose at first the network includes only one generator: since the network is connected, there exists certainly a path connecting points A and B and a path connecting points C and D (fig. 5). If the latter is entirely distinct from the former, there is a mesh including the generator, the branch b and both the paths. Let now all branches not included in such a mesh be cutted: then branch b will be traversed by a nonzero current. If on the other hand path A B has some part E in common with path CD (fig. 6) then consider all other paths between C and D: either all have in common with path AB the same part E, and then the network is hinged, which is contrary to our hypothesis; or some of them have another part E' in common with

AB, and then there exists a mesh such as AEDBE'CA, for which the above considerations can be repeated.

The proposition can be inverted, that is, if the network is not hinged, there is always a mesh including both the generator and the branch under discussion. If the generators are more than one, and hinges and series-connected ports are excluded, there is a mesh including branch b and generator 1, a mesh including branch b and generator 2 and so on. Eachone of these meshes, since no generators are series-connected, is at least in part distinct from any other. Let all branches be cutted but those pertinent to such meshes, and let the possible common parts among them be short-circuited: then the network looks like the one in fig. 7, in which it is evidently impossible that I_b vanishes identically.

Then we have proved that, if the n-port structure is not hinged nor contains series-connected ports, the branches cutting the previously said surface can be traversed by identically vanishing currents if and only if they are all internal and fed in the above said fashion.

It is to be now investigated whether such branches can be traversed by nonvanishing currents whose sum is however identically zero. It can be easily seen that if one at least of these branches is internal, this is impossible. In fact, for this internal branch, the current of which is by hypothesis nonvanishing, yields eq. (12) with $\sum V_r \neq 0$. Suppose now that in some configuration the currents of those branches fulfil eq. $\sum I_s = 0$; let the conductance of the said internal branch be varied: the current in it, by eq. (12), varies proportionally to it, while the voltage distribution among the nodes is left unchanged. Hence all other branches are still traversed by the same currents and the sum of the latter with the current in the considered internal branch will no longer be zero.

If the surface cuts only connecting branches besides ports i, let us consider a mesh including one of these branches b and a generator; such a mesh certainly exists, as proved above. Let all branches in this mesh be short-circuited, but the one including the generator and branch b; then the current in that branch is simply $I_b = g_b V$. But this case is exactly equal to the one previously examined, when one internal branch was admitted to cut the surface: hence in this case also it is impossible that the sum of the currents of the branches cutting the surface is identically zero.

We can resume the above results by the following Theorem 5: a necessary and sufficient condition in order that an n-port structure is affected by a linear rectangular degeneracy is that eitherone of the following conditions is fulfilled:

1) The n-port includes series-connected ports.
2) The network to which the n-port is reduced when all ports are short-circuited is hinged.
3) There exist a surface that, when some ports j are properly fed by voltage generators of ± 1 V, cuts only a) some ports i, b) short-circuited internal branches, c) internal branches not traversed by current because the voltage sum along the tree paths connecting their end nodes is vanishing.

Practically, the search for such degeneracies can be accomplished in the following way:

1) Look for series-connected ports.
2) If no series-connected ports, exist, short-circuit all ports and search for hinges in the so obtained network.
3) If no hinges are found, search for a closed surface cutting only ports and internal branches.
4) If such a surface exists and does not cut any port in at least one part of the graph of the ports, there are one or more linear degeneracies involving elements of a single column of the G_{ij} matrix: this can be seen by feeding any one of the ports in the previously said separate part.

5) If such a surface exists and cuts ports in all separate parts, one will examine eachone of the latter to ascertain if, by properly feeding one or more of its ports, its internal branches cutting the surface are all subjected to vanishing voltages.
6) If none of the above conditions is fulfilled, there is no linear rectangular degeneracy in the n-port.

We can prove now that the conditions of Theorem 5 are necessary for the existence not only of a rectangular degeneracy, but of a linear degeneracy at all. This means that any linear degeneracy necessarily implies some rectangular ones, and hence must be a linear consequence of them.

Any linear equation in the $G_{ij}^{(p)}$'s can be decomposed in many ways into a sum of rectangular terms. Consider at first the case in which such terms are exactly two:

$$\sum_{ij} G_{ij} = \sum_{h} \sum_{k} G_{hk} + \sum_{m} \sum_{n} G_{mn} = 0 \qquad (13)$$

If eachone of the two rectangular terms is identically vanishing, the theorem is proved. If such is not the case, denote as surface 1 the one cutting ports h, as surface 2 the one cutting ports m. Eq. (13) means that the sum of the currents in the branches cutting surface 1 when ports k are properly fed by ± 1 V generators and all other ports are short-circuited, identically coincides with the sum of the currents in the branches cutting surface 2 when ports n are properly fed by ± 1 V generators and all other ports are short-circuited. Remove all branches cutting surface 2 and driving current in the latter conditions; them the second term in the second member of (13) must vanish identically, and so must do, again because of (13), the first one. In this situation surface 1 must fulfil either condition 3 or condition 2 of Theorem 5 (condition 1 can be easily seen to be a particular case of condition 3). If condition 3 is fulfilled, replace one of the removed branches: if this did not cut surface 1, it would be impossible that the sum of the currents in the branches cutting surface 1 would be different from zero, since all these branches are subjected to identically vanishing voltages (condition 3); but since this is contrary to the hypothesis, such a branch must cut surface 1. By repeating this talk for all removed branches, we find that all must cut surface 1 and hence the situation looks like the one shown in Fig. 8. If condition 2 is fulfilled, this means that removing those branches induces an hinge between generators k and ports h. Repeat then the reasoning interchanging surface 1 with surface 2: either in this case no hinge is induced when branches cutting surface 1 are removed, and then the previous reasoning still applies: or the contrary happens, and then the network necessarily looks like the one in fig. 9; a careful examination of the figure shows that any other connection would infirm eq. (13). But fig. 9 can evidently be arranged as in fig. 10 which is an obvious generalization of fig. 8.

Let now ports k be fed by voltage generators in such a way that surface 1 is cutted by a total branch current I; the same current cuts also surface 2 (figg. 8 and 10). On the other hand, when ports n are fed by voltage generators in such a way as to give for the total branch current cutting surface 2, -I, the same current cuts also surface 1. Hence when both sets of ports are simultaneously excited, by the superposition principle, the current sum in the branches cutting surface 1 and in the branches cutting surface 2 both vanish. But this proves the existence of two rectangular degeneracies between the whole set of ports k and n and the two sets of ports h and m respectively.

It is to be noted that the reasonings which in fig. 10 apply to surfaces 1 and 2, apply equally well to a surface enclosing 1 and s and to surface 2, or to a surface enclosing 2 and s and to surface 1: therefore one can always reduce his considerations to fig. 8. Suppose now to have a linear degeneracy which splits into three rectangular terms:

$$\sum_{ij} G_{ij} = \sum_{h} \sum_{k} G_{hk} + \sum_{m} \sum_{n} G_{mn} + \sum_{p} \sum_{q} G_{pq} = 0 \qquad (14)$$

Let surfaces cutting ports _h_, _m_ and _p_ be denoted as surfaces 1, 2 and 3 respectively. Remove all branches cutting surface 1, which are traversed by current when ports _k_ are fed. Then the sum of the two last terms in eq. (14) must be identically equal to zero and hence surfaces 2 and 3 look like surfaces 1 and 2 in fig. 8. Repeat the reasoning interchanging surface 1 with surfaces 2 and 3 successively: then also the pairs 1 and 3, and 1 and 2 must look like the surfaces in fig. 8. But then surfaces 1, 2 and 3 are connected as shown in fig. 11. But here a surface enclosing 1, 2 , 3 cuts only ports and internal branches, identically subjected to zero voltage, when generators _k_, _n_ and _q_ are properly fed. Hence the condition 3 of Theorem 5 is fulfilled.

The proof can be easily extended to any number of rectangular terms.

Therefore we can state the following:

Theorem 6: a necessary and sufficient condition in order that an _n_-port structure is affected by a linear degeneracy is that one at least of the conditions of Theorem 5 is fulfilled.

5.- Nonlinear degeneracies in _n_-port structures with _p_=1

In an _n_-port where the graph of the ports is made of two separate parts, the port conductances are easily expressed in terms of the port conductances of the (n + 1)-port obtained by including the additional port in the working ones, as a particular case of fig. 1, by means of the following equation:

$$G_{ij}^{(1)} = G_{ij}^{(o)} - \frac{G_{in}^{(o)} \, G_{nj}^{(o)}}{G_{nn}^{(o)}} \qquad (15)$$

which is obviously a specialization of (1). Port n is here the one which is permanently open-circuited.

The fact that in this chapter nonlinear degeneracies are studied, whereas in n-ports with p = 0 any degeneracy is linear, does not contrast with Theorem 1 ; in fact, it happens in general that nonlinear degeneracies between $G_{ij}^{(p)}$'s imply analogous degeneracies between $G_{ij}^{(o)}$'s splitting in linear equations; but this does not mean that such a splitting happens also for $G_{ij}^{(p)}$'s.

It is obvious that any degeneracy depends exclusively on the lackness of branches, since a full network has been proved to be nondegenerate: hence it must be deduced exclusively from equations expressing the fact that some branches are not present. Express then the conductance of a branch connected between nodes a and b in terms of port conductances G_{ij}'s: for the previously said properties one obtains on equation of the type:

$$g_{ab} = \sum_i \sum_j G_{ij}^{(o)} \qquad (16)$$

where the sum in the second member fulfils the sign rule (par. 4). By substituting in (16) eqs. (15) one obtains:

$$g_{ab} = g'_{ab} + h_a h_b \qquad (17)$$

where

$$g'_{ab} = \sum_i \sum_j G_{ij}^{(1)}$$

and

$$h_a h_b = \sum_i \sum_j \frac{G_{in}^{(o)} G_{nj}^{(o)}}{G_{nn}^{(o)}} =$$

$$= \frac{1}{G_{nn}^{(o)}} \sum_i G_{in}^{(o)} \sum_j G_{nj}^{(o)}$$

Following that we have said, the equations on which we shall investigate degeneracy are of the type:

$$g'_{ab} + h_a h_b = 0 \qquad (18)$$

and are as many as the lacking branches.

It is well known (7) that the sum of h's extended to all nodes in the network is identically zero. Apart from this relation, h's can be either independent or linearly dependent, according to the fact that they are linear functions of the G'_{in}'s which in turn can fulfil eitherone of the said conditions.

In order to describe nonlinear degeneracies it is convenient to define the graph G of the lacking branches, where a branch is thought as lacking if it exists in the full network on the same nodes, but not on the given network. We study first

1) Nonlinear degeneracies when variables h are linearly independent. In such a case a functional relation between the g'_{ab} and hence between the $G^{(1)}_{ij}$'s can be due only to the structure of products $h_a h_b$. Let the graph G be considered and suppose it is made of g separate parts. Eachone of these includes at least a tree and then the following cases are to be considered:

1.1) Eachone of the g separate parts is just a tree. Then, if all g'_{ab}'s have been arbitrarily fixed, one is able to express all h's attached to nodes in that separate part as a rational function of one of them, which can be in turn arbitrarily fixed. If graph G does not include all nodes of the network, the independence of g'_{ab}'s is then assured. If such a graph includes all nodes, the sum of h's of all nodes must be still equated to zero. Then one can fix arbitrarily g-1 h's attached to g-1 nodes in different separate parts, and solve for the h relative to the last node of the last separate part a quadratic equation. Since the latter contains in its coefficients the arbitrary variables, one can always find a real value for the unknowns so that the network is still nondegenerate. However, if $g=1$, no h's can be fixed arbitrarily, so that the coefficients of the quadratic equation do contain only the g'_{ab}'s. For particular values of the latter, it can happen that the h's are complex conjugate, which means that no network exists, including positive or possibly negative resistors, which realizes the given matrix $G^{(1)}_{ij}$ with the prescribed absence of branches. We define such a structure as a quasi-degenerate one.

1.2) In at least one of the separate parts of graph G there is a chord. In order that the n-port structure is nondegenerate these chords cannot be more than $g + i - 1$, if i are the nodes not included in the graph G.

Moreover, in each separate part all odd chords, that is chords which form a mesh with an odd number of tree branches, yield a product $h_a h_b$ which is functionally dependent from the analogous products relative to tree branches (3). Hence, if an odd chord lacks, the corresponding g'_{ab} is functionally dependent on some other ones, and in general r lacking odd chords imply r functional relations between the g'_{ab}'s. On the other hand (3), in each separate part exactly one product $h_a h_b$ can be fixed which is associated with an even chord, that is with a chord forming a mesh with an even number of tree branches: it must be concluded that if in a separate part of G s even chords lack, this implies $s - 1$ functional relations between the g'_{ab}'s. It must be remarked that in such a case the branches which can lack without the network degenerating are exactly g even chords, and hence $i - 1$ branches potentially eliminable cannot be removed. If $i = 1$, all allowed branches can be removed. In this case the h's are determined by a quadratic equation yielding real or pure imaginary solutions. Hence there is no quasi-degeneracy.

If on the other hand graph G includes all nodes, then two are the cases; either it includes an odd chord or more than one even chord in any separate part and then the situation is just like the above one; or $g-1$ separate parts are trees plus an even chord and the latter is a tree (no other situation is possible, because otherwise too chords would lack). In the first $g-1$ parts all h's are determined by quadratic equations of the previously said type. In the last one we find a quadratic equation of general type and hence the network is quasi-degenerate.

2) Nonlinear degeneracies when variables $\underline{h}$ are linearly dependent. Suppose there is a functional relation between the $\underline{h}$'s due to the fact that they are linearly dependent. This must identically split into a sum of products of sums of $\underline{h}$'s, in each term of which one factor at least is identically zero. But such a relation implies a linear relation between the g'_{ab} and hence one is reported to the treatment of the previous paragraph.

6.- Conclusion

The aim of this work is to give simple rules for deciding whether an n-port network is degenerate or not, and eventually for writing down the equations of functional dependance between port conductances, without complex calculations, but utilizing direct researches on the topology of the network. The problem of finding linear degeneracies is completely solved; the nonlinear ones have been studied in the n-port networks with n + 2 nodes.

REFERENCES

(1) P. Slepian, in The Realization of n-Port Netowrk without Transformers - A Panel Discussion, IRE, Trans. on Circuit Theory, CT-9 (Sept., 1962), pp. 208-209.

(2) E.A. Guillemin, On the Analysis and Synthesis of Single-Element-Kind Networks, IRE Trans. on Circuit Theory, CT-7 (Sept., 1960), pp. 303-312.

(3) G. Biorci and P.P. Civalleri, On the Synthesis of Resistive n-Port Networks, IRE Trans. on Circuit Theory, CT-8 (March, 1961), pp. 22-28.

(4) G. Biorci and P.P. Civalleri, Conditions for the Realizability of a Conductance Matrix, IRE Trans. on Circuit Theory, CT-9 (Sept., 1961), pp. 312-317.

(5) See ref. (2), (3), (4).

(6) P.P. Civalleri, Topological Rules for Coordinate Transformations in Linear Reciprocal Networks, to be published in "Alta Frequenza".

(7) E.A. Guillemin, On the Realization of an nth-Order G Matrix, IRE Trans. on Circuit Theory, CT-8 (Sept., 1961), pp. 318-323.

(8) G. Biorci and P.P. Civalleri, About a Basic Theorem on Resistive Networks, presented at the 1963 IRE International Convention.

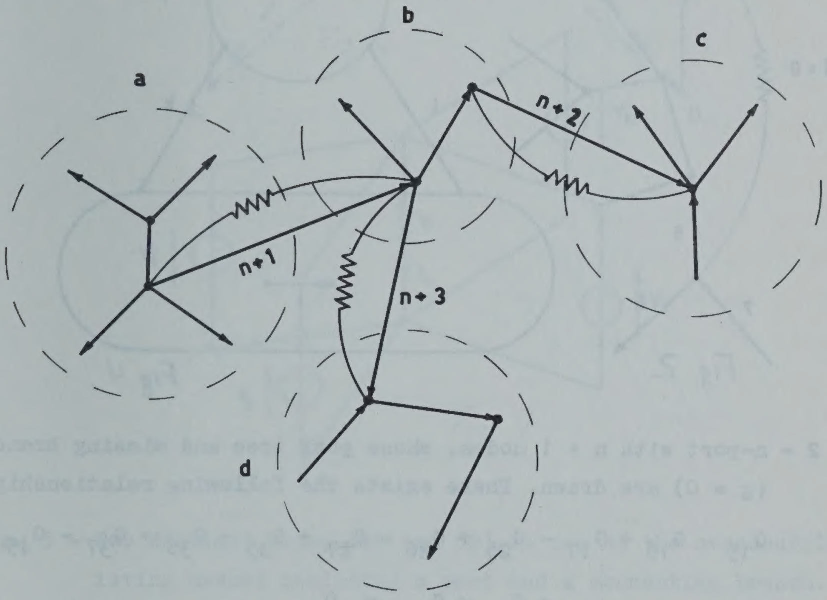


Fig. 1 - Network in which the separate parts a, b, c, d of the graph of the ports are connected by additional ports n + 1, n + 2, n + 3 arranged across nodes connected by actual branches.

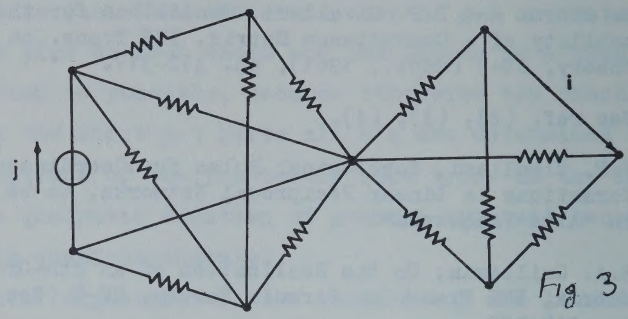


Fig 3

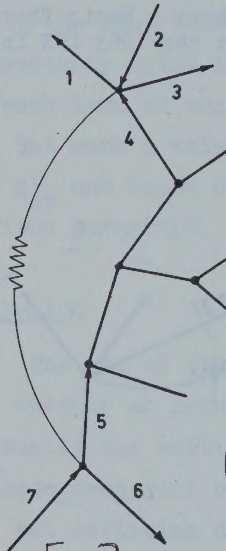


Fig 2

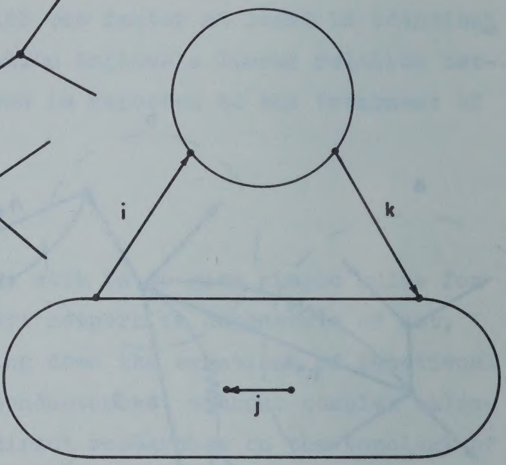


Fig 4

Fig. 2 - n-port with n + 1 nodes, whose port tree and missing branch (g = 0) are drawn. There exists the following relationship:

$$G_{15} - G_{16} + G_{17} - G_{25} + G_{26} - G_{27} + G_{35} - G_{36} + G_{37} - G_{45} + G_{46} - G_{47} = 0$$

Fig. 3 - Rectangular degeneracy induced from the fact that short-circuiting all ports yield a hinged network.

Fig. 4 - Rectangular degeneracy induced from the presence of ports in series.

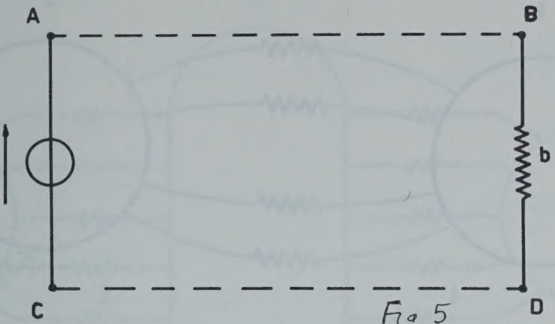


Fig 5

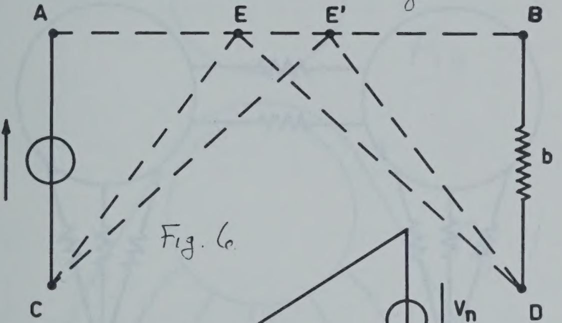


Fig. 6

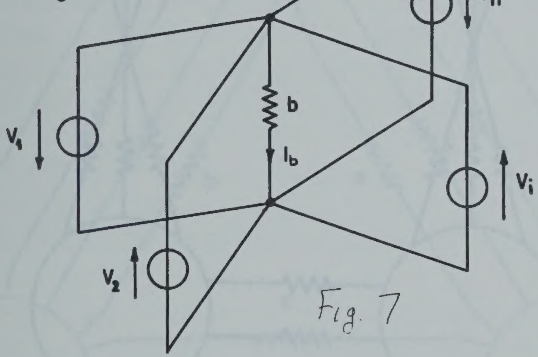
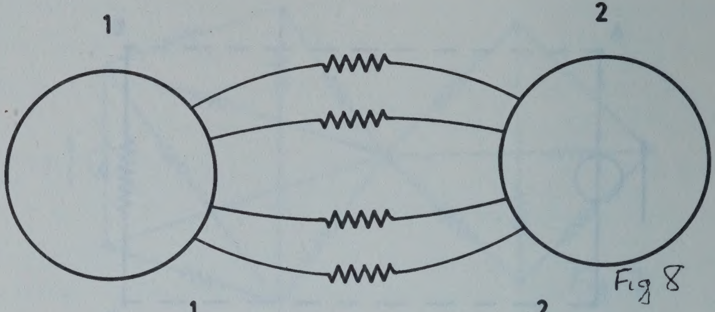


Fig. 7

Fig. 5 - Two distinct paths (AB and CD) in one of the necessarily existing meshes including a port and a connecting branch.

Fig. 6 - Case in which the two paths AB and CD have a part in common.

Fig. 7 - Meshes including a connecting branch b and eachone of the ports.



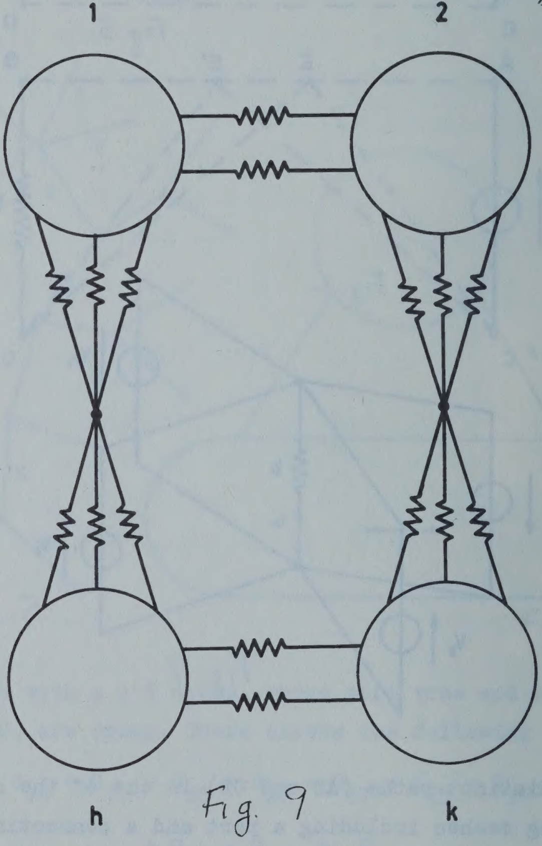


Fig. 8 - Surfaces 1 and 2 with branches traversed by currents.

Fig. 9 - Structure of the network of fig. 8 when suppressing branches traversed by currents in either surface yields an hinge.

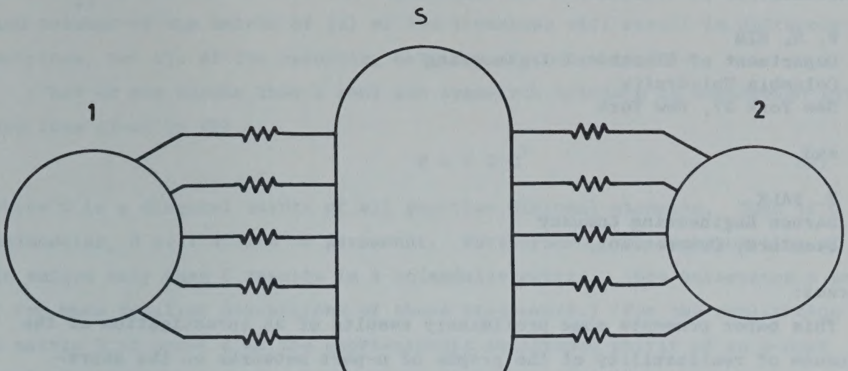


Fig. 10

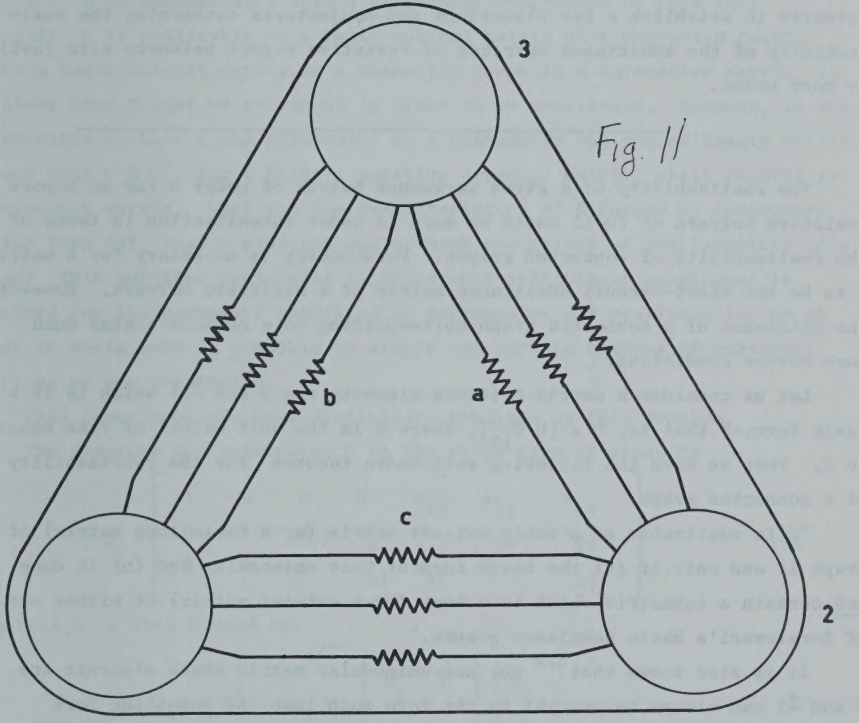


Fig. 11

Fig. 10 - Simplified representation of fig. 9.

Fig. 11 - Surfaces 1, 2 and 3 with branches traversed by currents.

CHARACTERISTICS OF NONPLANAR N-PORT NETWORKS*

W. H. KIM
Department of Electrical Engineering
Columbia University
New York 27, New York

and

T. FALK
Barnes Engineering Company
Stamford, Connecticut

Abstract:

This paper presents some preliminary results of an investigation of the influence of realizability of the graphs of n-port networks on the short-circuit admittance matrices of the networks. Some known results of the study of the theory of linear graphs are combined with the realization of n-port networks to establish a few assertions and conjectures concerning the realizability of the admittance matrices of resistive n-port networks with (n+1) or more nodes.

The realizability of a given paramount matrix of order n for an n-port resistive network of (n+1) nodes or more is under investigation in terms of the realizability of connected graphs. Paramountcy is necessary for a matrix H to be the short-circuit admittance matrix of a resistive network. However, the existence of a connected graph corresponding to a network claims much more severe conditions.

Let us consider a matrix C (whose elements are 0 and ±1) which is in a basic form;** that is, $C = [U \; C_{12}]$, where U is the unit matrix of rank equal to C. Then we have the following well-known theorem[1] for the realizability of a connected graph.

"C is realizable as a basic cut-set matrix (or a basic loop matrix) of a graph if and only if (a) the basic form of C is unimodular and (b) it does not contain a submatrix which is a loop (or a cut-set matrix) of either one of Kuratowski's basic nonplanar graphs."

It is also known that[1,2] any non-unimodular matrix whose elements are 0 and ±1 can always be brought to the form such that the resulting form contains the matrix of the form (1) or its transpose.

$$T = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} \\ a_{21} & 0 & a_{23} & a_{24} \\ a_{31} & a_{32} & 0 & a_{34} \end{bmatrix} \tag{1}$$

---

*This work is partly supported by National Science Foundation Grant GP-630 and by the Barnes Engineering Company.

**The terminology used in this paper is consistent with that in Reference (5).

where a_{ij} can be either o or -1. Note here that rearrangement of the rows and columns of the matrix of (1) or its transpose will result in different matrices, but all of the resulting matrices are non-unimodular.

Let us now assume that a real and symmetric matrix H is decomposed into the form given in (2).

$$H = C\ D\ C^t \tag{2}$$

where D is a diagonal matrix of all positive diagonal elements. Then if C is unimodular, H will always be paramount. Furthermore, the decomposition of H is unique only when C results in a unimodular matrix. (See References 5 and 7 for more detailed discussions of these statements.) For the realization of a matrix H of order n as the short-circuit admittance matrix of an n-port resistive network with (n+1) nodes the following conditions are then necessary and sufficient:

(i) H is decomposable into a unimodular congruence of form (2);

(ii) C is realizable as a basic cut-set matrix of a connected graph.

Since a basic cut-set matrix of a connected graph is a unimodular matrix, it is clear that H must be paramount in order to be realizable. However, it may be possible to have a non-unimodular or a unimodular but nonrealizable matrix C such that $C\ D\ C^t$, for D being a positive diagonal matrix, still results in a paramount matrix. What are the characteristics of H formed by congruence of the form (2), when C violates one or both conditions of realizability of a graph? This question would shed light on what sufficiency complement is required for the necessary condition of paramountcy for realizability or at least it would make it possible to single out certain classes of paramount matrices as not realizable.

This paper presents some preliminary findings on this problem.

The simplest non-unimodular C in the basic form is given by

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & a_{12} & a_{13} & a_{14} \\ 0 & 1 & 0 & a_{21} & 0 & a_{23} & a_{24} \\ 0 & 0 & 1 & a_{31} & a_{32} & 0 & a_{34} \end{bmatrix} \tag{3}$$

A matrix H is then formed as:

$$H = C\ D\ C^t = [U_3\ \ T] \begin{bmatrix} D_3 & 0 \\ 0 & D_4 \end{bmatrix} \begin{bmatrix} U_3 \\ T^t \end{bmatrix} \tag{4}$$

where U_k is a unit matrix of order k and D_k is a diagonal matrix with positive diagonal elements of order k.

Evaluating (4) and considering that $a_{ij}^2 = 1$ for $i \neq j$, it can readily be shown by direct computation that this matrix is paramount for $a_{ij} = 1$ if all diagonal elements d_k are positive. "For all positive D the non-unimodular congruence (4) is paramount for $a_{ij} = 1$."

For the general case an argument of plausibility shall be developed to show that congruences formed by matrices with nonunimodular submatrices and diagonal matrices with non-negative elements are paramount.

Reordering rows of the matrices in the congruent product will not change any conclusions as to paramountcy. It is justified then to consider a matrix C of order n by k, as

$$C = \begin{bmatrix} A_1 & A_3 \\ A_2 & T \end{bmatrix} \tag{5}$$

where T is nonunimodular but A_1, A_2, and A_3 are unimodular and $c_{ij} = 0,1$.

Consider the congruence $H = C\ D\ C^t$ with C of the form (5) and D a diagonal matrix of order k with nonegative elements. Partitioning D,

$$D = \begin{bmatrix} D_{k-4} & 0 \\ 0 & D_4 \end{bmatrix} \tag{6}$$

and performing the congruent product will yield

$$H = C\ D\ C^t = \begin{bmatrix} A_1\ D_{k-4}\ A_1^t + A_3\ D_4\ A_3^t & A_1\ D_{k-4}\ A_2^t + A_3\ D_4\ T^t \\ A_2\ D_{k-4}\ A_1^t + T\,D_4\ A_3^t & A_2\ D_{k-4}\ A_2^t + T\ D_4\ T^t \end{bmatrix} \tag{7}$$

Examination of the matrices in the main diagonal of the partitioning of H shows that they are paramount. This is due to Cederbaum's result[2] that a unimodular congruence of a diagonal matrix with nonegative elements is paramount and TD_4T^t is also paramount from the previous discussion (Eq. 4). If the place of T were taken by a unimodular matrix, then H would certainly be paramount. Thus the role of $A_1\ D_{k-4}\ A_2^t$ does not have to be investigated. It remains to examine $A_3\ D_4\ T^t$.

Assume A_3 to be of its simplest form:

$$A_3 = [\alpha_1\ \ \alpha_2\ \ \alpha_3\ \ \alpha_4] \tag{8}$$

with $_i = 0, 1$. We now form

(9)

$$D_4 = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_3 \end{bmatrix}$$

Also obtain

$$A_3\ D_4\ T^t =$$

$$[\alpha_2 a_{21} d_2 + \alpha_3 a_{31} d_3 + \alpha_4 a_{41} d_4 \quad \alpha_1 a_{12} d_1 + \alpha_3 a_{32} d_3 + \alpha_4 a_{42} d_4 \quad \alpha_1 a_{13} d_1 + \alpha_2 a_{23} d_2 + \alpha_4 a_{43} d_4] \tag{10}$$

Thus the contribution of A_3 in its assumed form (8) for the main diagonal element $(A_3\ D_4\ A_3^t)$ is greater than or equal to the absolute value of its contribution to any of the off-diagonal elements (elements of the single row matrix $A_3\ D_4\ T^t$). If A_3 consists of several rows, this relationship will still hold.

The above argument indicates that the paramountcy of a unimodular congruence would hold in a stronger form:

"The matrix H formed by the congruence $H = C\ D\ C^t$ is paramount if the elements of the diagonal matrix D are nonnegative and the elements of C are 0, +1".

We shall next investigate the presence of cut-set (or loop) matrices of Kuratowski's basic non-planar graphs (i.e., violation of the second condition of the realizability of graphs). The basic loop matrices of non-planar graphs, pentagon and hexagon, K_p and K_h, are:

$$K_p = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} = [U_b\ K_{po}] \tag{11}$$

$$K_h = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} = [U_4\ K_{ho}] \tag{12}$$

These matrices,[3] being unimodular, will produce paramount congruent products.

With the basic loop matrix K_p corresponding to the nonplanar graph of Fig. 1, we form the congruent product $K_p\ D_{10}\ K_p^T$, where

$$D_{10} = \begin{bmatrix} a & & & & & & & & & \\ & b & & & & & & & 0 & \\ & & c & & & & & & & \\ & & & d & & & & & & \\ & & & & e & & & & & \\ & & & & & f & & & & \\ & & & & & & g & & & \\ & & & & & & & h & & \\ & 0 & & & & & & & i & \\ & & & & & & & & & j \end{bmatrix} \tag{13}$$

[Fig. 1.]

It is seen that a, b, c, d, e and f are the chords corresponding to the tree g, h, i and j on which K_p is based. Then,

$$K_p D_{10} K_p^t = [U_6 \ \ K_{po}] \ D_{10} \begin{bmatrix} U_6 \\ K_{po} \end{bmatrix} = G_{kp} \tag{14}$$

where K_{po} is the submatrix of K_p consisting of the last four columns. Thus we have

$$G_{kp} = \begin{bmatrix} a & & & & & 0 \\ & b & & & & \\ & & c & & & \\ & & & d & & \\ & & & & e & \\ 0 & & & & & f \end{bmatrix}$$

$$+ \begin{bmatrix} g+h+i+j & g+i & g+i+j & g+h+j & g+j & h+j \\ g+i & g+i & g+i & g & g & 0 \\ g+i+j & g+i & g+i+j & g+j & g+j & j \\ g+h+j & g & g+j & g+h+j & g+j & h+j \\ g+j & g & g+j & g+j & g+j & j \\ h+j & 0 & j & h+j & j & h+j \end{bmatrix} \tag{15a}$$

or

$$G_{kp} = D_6 + G_{kpo} \tag{15b}$$

and

$$G_{kpo} \doteq K_{po} D_4 K_{po}^t \tag{15c}$$

where D_6 and G_{kpo} are the first and second matrices of (15a), respectively.

One particular assumption for the orientation of the edges of the graph in Fig. 1 has resulted in all positive or zero elements in K_{po} and in $K_{po} D_4 K_{po}^t$. (The elements of D_{10} are nonnegative.) Other possible orientations can introduce negative signs in G_{kpo}.

In the following it will be shown that matrices of the structure of G_{kpo} or rearrangements of G_{kpo} cannot be realized with any number of nodes as admittance matrices of 6-port graphs. (This of course would also be true of any (n x m) matrix that contains a principal submatrix of order six of such structure.)

G_{kpo} cannot be realized with (n+1 = 7) nodes. It has been generated by a unimodular congruence using the non-planar pentagon's basic loop matrix which is not realizable as a cut-set matrix needed here. Also, because the congruence is unimodular it can be considered as a unique decomposition not leaving any possibility of a seven-node realization.

Further inspection of G_{kpo} reveals that it exhibits the property (and of course with any combination of signs in the off-diagonal elements)

$$|y_{1K}| = y_{KK} \quad \text{for} \quad K = 1, 2, 3, 4, 5, 6 \tag{16}$$

A previous result[4] indicates that an (n x n) matrix that exhibits property (16) can either be realized with n+1 nodes or cannot be realized at all as a transformerless, passive network. In this case it cannot be realized with n+1 = 7 nodes; thus it cannot be realized at all.

To continue the argument, perform a formal realization of G_{kpo} with all positive elements, although inspection of the all-positive matrix with its lack of taper even after rearranging would indicate the impossibility of realization with (n+1) nodes even if it were not shown that a nonplanar loop matrix decomposition existed. Using the procedure outlined[5] G_{kpo} is realized with a linear-tree port-structure and the following edge-conductance matrix:

$$G_e = \begin{bmatrix} 0 & 0 & i & 0 & g & 0 \\ 0 & 0 & 0 & 0 & 0 & j \\ i & 0 & h & -h & 0 & h \\ 0 & 0 & -h & h & 0 & -h \\ g & 0 & 0 & 0 & 0 & h \\ 0 & j & h & -h & h & 0 \end{bmatrix} \tag{17}$$

A six-port network corresponding to (17) is shown in Fig. 2. It is seen that it can only be realized with two negative conductances if $h > 0$.

[Fig. 2]

As to the realizability of a matrix of the structure (or rearrangement thereof) of $G_{kp} = D_6 + G_{kpo}$ with non-zero elements in D_6 the only statement that can be made is that it is not realizable with seven nodes, but nothing can be said yet on realizability with a larger number of nodes as condition (16) is here not satisfied. Augmentation of the matrix[6] or similar procedures should be used to decide upon more than 7-node realizability. If negative elements are allowed, it can be said (for the all positive element case) that if a matrix of order six is the sum of a uniformly tapered matrix and a Kuratowski pentagon congruence then it can be realized with a linear port-structure and at most two negative conductances. Statements of similar character could be made on matrices with not all positive elements.

In the general case where K_p is embedded in a larger matrix C of order n by Γ,

$$C = \left[\begin{array}{cccc} & C_{11} & C_{12} & C_{13} \\ U_n & C_{21} & K_p & C_{23} \\ & C_{31} & C_{32} & C_{33} \end{array}\right] n \tag{18}$$

with C_{ij} realizable cut-set matrices the congruence with a diagonal matrix of order r results in an (n x n) matrix that contains only one submatrix of order six that is the sum of a matrix realizable with seven nodes and matrix of the form of G_{kpo} of (15c). Then the above considerations regarding realization with at most two negative conductances will hold.

As to means to recognize the similarity of a given matris of order six (or submatrix) to G_{kpo} or G_{kp} several programmable procedures could be devised.

Further work should produce procedures to extract G_{kpo} from its parallel combination with a realizable matrix in some minimal sense.

References

1. Tutte, W.T.: "Matroids and Graphs," Trans.Am.Math.Soc. 90, 1959, 527-552.

2. Cederbaum, I.: "Matrices All of Whose Elements and Subdeterminants are 1, -1 or 0," J.Math. and Phys. 36(1958) 351-361.

3. Belevitch: On the Realizability of Graphs with Prescribed Circuit Matrices, Lockheed Symposium on Switching Theory, 1962.

4. Cederbaum, I.: Topological Considerations in the Realization of Resistive n-port Networks, IRE, PGCT Trans. Vol. 9, Sept., 1961; pp.324-329.

5. Kim, W.H. and Chien, R.T.: Topological Analysis and Synthesis of Communication Networks, Columbia Press, 1962.

6. Guillemin, E.A.: On the Realization of an n-th Order G-Matrix, IRE, PGCT Trans., CT-9, Sept., 1961, pp.318-323.

7. Cederbaum, I.: "Applications of Matrix Algebra to Network Theory" IRE, PGCT Trans. Special Supplement, CT-6, May 1959, pp.127-137.

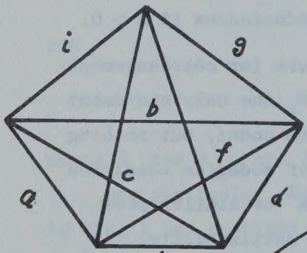


Fig. 1

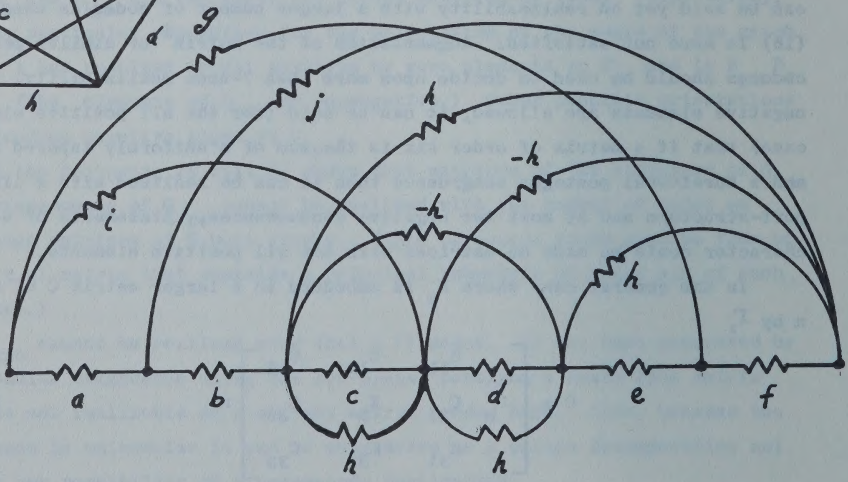


Fig. 2

ON THE STRUCTURE OF BASIC AMBIT SETS

H. E. MEADOWS
Columbia University
New York, New York

and

B. LIU
Princeton University
Princeton, New Jersey

The analysis and synthesis of various kinds of lumped networks, such as n-port resistive networks, communication nets and time-variable networks, depend upon the ability to relate internal constraints imposed by the elements and their connections to the external behaviour at chosen ports. Regarding the topological constraints imposed by the connections in the examples cited, we may observe that the concept of a tree usually is fundamental to both analysis and synthesis. Recently, several participants in a panel discussion[1] emphasized a need for novel analytical approaches to network problems in the realization of n-port networks without transformers. In this paper the concept of adjacency of ambits is employed to make what, hopefully, may be a step toward a useful and somewhat different characterization of lumped networks[2]. Primary attention is given to the description of a tree provided by the structure of the basic ambits associated with its branches[3]. The potential application of such an alternate description lies primarily in its relation to the so-called "port structure" of an n-port network; however, the similarities between electrical networks and communication nets[4] also help to motivate the following discussion.

As indicated by Dasher[5], the statement of Kirchhoff's current law in terms of a closed surface, and the intersection of such a surface with the flat surface on which a network is often drawn, naturally lead to the concept of an ambit, a concept essentially dual to that of a loop. Intuitively, we may say that a loop is a circuit roughly parallel or tangent to a set of edges of a graph or network, while an ambit is a circuit intersecting a set of edges. Dasher showed the utility of the ambit concept in network analysis.

An application of the ambit concept to network synthesis was given by Okada and others[6,7]. These workers recognized clearly the application of Jordan's theorem[8] and utilized the fact that basic ambits corresponding to a tree may be drawn without intersections. These considerations are also fundamental to the present paper.

The following summary outlines the organization of this paper: To provide a basis for further discussion, the concept, definition and properties of ambits are reviewed and the concept of a basic ambit set is introduced. Then a new graph, called an ambit adjacency graph, is described, and necessary and sufficient conditions for a graph to be an ambit adjacency graph are given. The ambit adjacency graph or matrix of a basic ambit set is used to represent a tree; thus, an alternate characterization and a simple procedure for realizing a tree arise from the ambit adjacency graph. Distances between ambits are defined and a simple and efficient means to compute the distance matrix is given. Finally, the close relation between these distances and the structure of a tree or its basic ambit set is illustrated. In general, proofs of properties are omitted for brevity throughout.

Definition and Convention

In this paper, the linear graph terminology of Kim and Chien[9] will be followed generally, and we shall represent networks by linear graphs "drawn" (with crossed edges and nodes distinguished) on a plane (or, in certain cases, a sphere). Any other entities, including ambits, also lie in the same plane; thus, we may define an ambit as follows:

Definition 1. An ambit of a linear graph is a closed Jordan curve[8] which does not intersect nodes of a graph.

Since the graph consists of nodes and edges, an ambit may intersect edges of the graph, i.e., an ambit may have one or more points in common with a given edge. We impose the requirement that the number of such points be at most one for each edge, and we do not permit tangency of ambits and edges. These restrictions may require deformation of edges. We shall also require for convenience that any ambit must intersect at least one edge. Some ambit are shown by broken lines on the graph of Fig. 1 for illustration.

Properties of Ambits

Three important properties of ambits are direct consequences of their definition, Kirchhoff's current law, and Jordan's theorem, which states that "A closed Jordan curve divides the plane into two domains having the Jordan curve for their common boundary[8]." It is convenient to note that one of these domains is bounded and the other unbounded; thus, we may identify the inside and outside, respectively, for an ambit.

Property 1. Each ambit separates the nodes of the graph into two disjoint sets, namely the nodes inside and outside of the ambit. In view of the previous restriction on ambits to be considered here, neither of these sets is empty.

Property 2. The set of edges intersecting an ambit comprises a seg[10], i.e., a cut set or a union of cut sets.

Property 3. The sum of the edge currents entering any ambit is zero.

In this paper our attention is centered on a particular class of ambits to be called basic ambits. These ambits are uniquely associated with the branches of a tree of the graph.

Definition 2. A basic ambit corresponding to a particular branch of a tree is an ambit having one intersection with that branch and no intersections with other branches of the tree.

Definition 3. A basic ambit set associated with a tree is a set of basic ambits each corresponding uniquely to a distinct tree branch; or,

Definition 3'. A basic ambit set associated with a tree of a graph is a maximal set of ambits such that each branch of the tree is intersected once by the ambits in the set. A tree and its associated basic ambit set are illustrated in Fig. 2(a). An important property of basic ambits is:

Property 4. Associated with each tree, there exists a non-intersecting basic ambit set. This set is determined uniquely by the tree. Here we regard two basic ambit sets as equivalent if one set may be homeomorphically deformed to coincide with the other without encountering intersections between ambits. In order to allow inversion of ambits (interchange of inside and outside), we may regard the plane on which the graph is "drawn" as mapped on a sphere; inversions of ambits are then obtained by deformations in which the ambits are allowed to pass through the image of the point at infinity.

A basic relation between two ambits is that of adjacency; this relation is fundamentally important throughout this paper.

Definition 4. Given a set of non-intersecting ambits on a graph, two ambits are adjacent if an edge may be added between two nodes of the graph to intersect these two ambits and no others. In the absence of a graph, two of a set of non-intersecting ambits are adjacent if there exists a path in the plane which intersects these two ambits and no other.

We shall use this definition to define an ambit adjacency graph and matrix in the next section. However, first we state an important property of basic ambit sets based on Jordan's theorem and the preceding definition.

Property 5. Any basic ambit separates the other basic ambits of the same tree into two sets A and B such that no ambit in A is adjacent to any ambit in B.

The Ambit Adjacency Graph and Matrix

Consider an n-branch tree and its basic ambit set. The uniquely associated ambit adjacency graph is an n-node linear graph in which each node represents a basic ambit and two nodes are joined by an edge if an only if the corresponding two ambits are adjacent. For example, the ambit adjacency graph for the basic ambit set of Fig. 2(a) is shown in Fig. 2(b). We shall not distinguish between "ambit i" and "the node in the ambit adjacency graph corresponding to ambit i" since the meaning will always be clear from the context. In order to give the necessary and sufficient conditions for a graph to be an ambit adjacency graph, we utilize the concepts of completely connected subgraph and articulation point, for which we employ the following standard definitions.

Definition 5. A completely connected subgraph is a subgraph such that precisely one edge joins each pair of nodes.

Definition 6. An articulation point (cut vertex or cut node) of a graph is a node a which separates the other nodes of the graph into two non-empty sets B and C such that any path linking a node of B and a node of C must pass through node *a*.

We shall have use for a particular kind of articulation point, which shall be called a two-articulation point, defined as follows:

Definition 7. *A two-articulation point* is an articulation point *a* such that sets B and C are unique except for a possible interchange of B and C.

The branches of a tree may be divided into two distinct classes: tip branches and non-tip branches[11]. In view of property 5, the basic ambit cutting a non-tip branch corresponds to an articulation point in the ambit adjacency graph. Suppose each non-tip branch of a tree is bisected into two separate (unconnected) branches; then the graph is separated into a set of star trees. Since the basic ambits associated with a star tree are all adjacent, we are led to the following necessary and sufficient condition for a graph to be the ambit adjacency graph corresponding to a tree. (The sufficiency is proved in the next section.)

Theorem 1. The necessary and sufficient conditions for a graph to be the ambit adjacency graph of a basic ambit set are:

(1) The graph is connected.

(2) The subgraphs obtained by bisecting all two-articulation points are completely connected. (The nature of the bisection is clearly implied above.)

A useful tool in the study of ambits is the ambit adjacency matrix.

Definition 8. The ambit adjacency matrix of a set of n non-intersecting ambits, (viz. a basic ambit set) is a symmetric n x n matrix A, of which each diagonal element a_{ii} is zero and each off-diagonal element a_{ij} is 1 or 0 according as ambits i and j are or are not adjacent, respectively.

The ambit adjacency matrix of the basic ambit set of figure 2(a) is

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \qquad (1)$$

It is important to note here that the ambit adjacency graph and matrix of a set of non-intersecting ambits bear a one-to-one correspondence.

For a star tree and a chain (linear) tree respectively, the matrices have the special forms (see Fig. 3),

$$A_{star} = \begin{bmatrix} 0 & 1 & 1 & \ldots & 1 \\ 1 & 0 & 1 & \ldots & 1 \\ 1 & 1 & 0 & \ldots & 1 \\ . & & & & \\ . & & & & \\ . & & & & \\ 1 & 1 & 1 & \ldots & 0 \end{bmatrix} \qquad (2)$$

and

$$A_{chain} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ 0 & 1 & 0 & .. & 0 \\ . & & & & \\ . & & & & \\ . & & & & \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} \qquad (3)$$

We shall see later that the "dispersions" of these two trees possess extremal values.

The Ambit Adjacency Graph as a Representation of a Tree

The importance of the basic ambit set of a tree and its ambit adjacency graph lies not only in the fact (property 4) that the basic ambit set and its adjacency graph or matrix are uniquely determined by the tree but also in the inverse statement that an ambit adjacency graph or matrix determines unqiuely the tree from which the basic ambit set is constructed. Thus, we may regard an ambit adjacency matrix as a means to represent a tree without resort to node-edge incidence matrices. Together with Theorem 1, the foregoing observations furnish necessary and sufficient conditions for a symmetric matrix of order n to represent a tree. These conditions, expressed as they are in terms not involving detection of loops, may prove useful in topological network analysis and synthesis, since they provide a different insight into the nature of a tree in relation to its basic ambit sets. To demonstrate the validity of these conditions, we now give a procedure to construct the basic ambit set and its corresponding tree from the ambit adjacency graph or matrix (thus completing the proof of Theorem 1).

Construction Procedure:

(1) Choose arbitrarily any initial ambit A_1 and draw this ambit as a simple closed curve.

(2) Select any ambit A_2 which is adjacent to A_1; since the ambit adjacency graph is connected there is such an A_2.
Draw this ambit as a simple closed curve lying outside of A_1.
(Placing A_2 outside of A_1 is an arbitrary choice; however, the location of the rest of the ambits with respect to inside and outside will be

fixed.)

Steps (1) and (2) initiate the construction, the process of which will now be described in inductive terms.

(3) Select any ambit A_i adjacent to an ambit A_j previously drawn and perform inversions, if necessary, so that A_j lies inside of no ambit previously drawn. The new ambit A_i cannot lie inside any ambit which is inside of A_j or inside any ambit which is outside of A_j because of Def. 4. Thus, in order to avoid intersection of ambits and to obtain adjacency of A_i and A_j, ambit A_i can only lie in one of two possible domains, viz., the intersection of the inside of A_j or the intersection of the outside of A_j and the outside of those ambits already outside of A_j. Because of property 5, A_i cannot be adjacent simultaneously to two ambits one inside and one outside of A_j. Thus, location of A_i is fixed by its adjacency or non-adjacency to any other previously drawn ambit adjacent to A_j.

(4) Repeated application of step (3) will exhaust the ambits not yet drawn.

We may observe at this point that we have realized by a finite construction the basic ambit set. We also have proved the sufficiency of the conditions in Theorem 1 because A_i must be adjacent to all ambits which are part of the boundary of the domain in which it must be placed. Now we proceed to construct the tree corresponding to the basic ambit set.

(5) Because of Jordan's theorem, the intersection of the exteriors and/or the interiors and exteriors of adjacent ambits form a number of domains one greater than the number of non-intersecting ambits. Exactly one node of the basic tree must be placed in each of these domains, which have been referred to as terraces in the literature.[7] There must be exactly n + 1 nodes for n basic ambits, and none of the terraces can be void of nodes since each tree branch must intersect only one basic ambit.

(6) Corresponding to each basic ambit a tree branch is drawn to intersect only that ambit. The terminals of such a branch are the nodes placed in thetwo domains (of the n + 1 domains) of which the chosen ambit consists of boundary points. Repeating this step for all of the basic ambits completes construction of the tree.

Distances between Ambits

By using the ambit adjacency graph, we may define a distance between the ambits of a basic ambit set (or, more generally, of a non-intersecting set of ambits).

Definition 9. The distance between ambits i and j is the minimum number of edges in a path connecting nodes i and j in the ambit adjacency graph; for example, in Fig. 2, the distance between ambits 1 and 2 is 1, between ambits

1 and 4 is 2 and between 2 and 5 is 3. We may define a distance matrix D for the basic ambit set by:

Definition 10. The distance matrix D is a square matrix with diagonal elements d_{ii} equal to zero and d_{ij} equal to the distance between ambits i and j.

It has been shown[12] that the distance matrix D can be derived from the matrix A by the following theorem:

Theorem 2. There exists an integer k such that $(A')^k = (A)^{k+1}$, where A' is the matrix derived from A by replacing the off-diagonal zeros by ∞'s and where matrix multiplication employs the following arithmetic:

$$xy = x + y, \quad \infty \cdot x = x \cdot \infty = \infty \tag{4}$$

$$x + y = \min(x,y) \text{ for all } x \text{ and } y. \tag{5}$$

The distance matrix D is then $D = (A')^k$. An especially simple calculation procedure based on this theorem is given in the Appendix.

The distance matrix is related closely to the structure of the tree, as is illustrated by the following property:

Property 6. If the entry d_{ij} in D is no less than any other entry in D, then ambits i and j correspond to two tips of the associated tree.

A measure which appears to be useful in describing the gross structure of a basic ambit set is its dispersion[12], defined in terms of the elements in the distance matrix as follows:

Definition 11. The dispersion of an n-ambit basic ambit set is the quantity

$$\Delta = \sum_{i,j=1}^{n} d_{ij}, \tag{6}$$

where d_{ij} is the distance between ambits i and j.

The distance matrix of the ambit adjacency graph associated with an n-branch star tree is

$$D_{star} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 1 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ 1 & 1 & 1 & \cdots & 0 \end{bmatrix} \tag{7}$$

It is easily seen that the dispersion $\Delta_{star} = n(n - 1)$ is minimum among all n-ambit basic ambit sets. The distance matrix of the basic ambits of an n-branch chain tree is, after a possible reordering,

$$D_{chain} = \begin{bmatrix} 0 & 1 & 2 & \cdots & n-1 \\ 1 & 0 & 1 & \cdots & n-2 \\ 2 & 1 & 0 & \cdots & n-3 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ n-1 & n-2 & n-3 & \cdots & 0 \end{bmatrix} \tag{8}$$

It may be shown that the dispersion in this case,

$$\Delta_{chain} = \frac{n(n^2 - 1)}{3}, \tag{9}$$

is the maximum attainable by the ambit adjacency graph associated with any n-branch tree.

Appendix

The appendix presents an especially simple procedure to calculate the distance matrix D from A, using the algebra introduced by Shimbel[12]. We first show that the matrix multiplication using the special arithmetic employed in Theorem 2 is associative, i.e. we show that

$$(XY)Z = X(YZ) \quad \text{where } X = (x_{ij}),\ Y = (y_{ij}), \text{ and } Z = (z_{ij}).$$

This is seen to be trivially true by noting that the ijth element in XY is $\underset{k}{\mathrm{Min}}(x_{ik} + y_{kj})$. Thus the ijth element in (XY)Z is

$$\underset{l}{\mathrm{Min}}\ [\underset{k}{\mathrm{Min}}\ (x_{ik} + y_{kl}) + z_{lj}] = \underset{k,l}{\mathrm{Min}}\ [x_{ik} + y_{kl} + z_{lj}].$$

Similarly, the ijth element of X(YZ) is

$$\underset{l}{\mathrm{Min}}\ [x_{il} + \underset{k}{\mathrm{Min}}\ (y_{lk} + z_{kj})] = \underset{k,l}{\mathrm{Min}}\ [x_{il} + y_{lk} + z_{kj}]$$

Interchanging the indices k and l establishes the assertion that (XY)Z = X(YZ). It therefore follows from a theorem in algebra[13] that $(A')^{k_1} (A')^{k_2} = (A')^{k_1+k_2}$. Thus to calculate D from A', one only need calculate successively the matrices $(A')^{2^m}$, m = 1,2,3, ..., using the fact that

$$(A')^{2^m} = (A')^{k_0} \quad \text{if } 2^m > k_0.$$

Using this procedure, the number of calculations to obtain D from A' is m_0 where m_0 is the minimum integer satisfying $2^{(m_0-1)} \geq k_0$. For any reasonable k_0, m_0 is clearly much less than k_0, the number of calculations required if successive powers of A' are calculated. Moreover, in our case we may always choose $2^{(m_0-1)} \geq n - 1 \geq k_0$ for simplicity.

Notes and References

1. The Realization of n-Port Networks Without Transformers--A Panel Discussion," IRE Trans. on Circuit Theory, vol. CT-9, pp. 202-214; September, 1962.
2. For brevity, we refer to the bibliography of reference 1 to indicate most of the background for this paper.
3. It is hoped that a somewhat more general network characterization related to the present characterization of a tree may be described in a later paper.
4. I. T. Frisch and W. H. Kim, "N-Port Resistive Networks and Communication Nets," IRE Trans. on Circuit Theory, vol. CT-8, pp. 118-120; December, 1961.
5. B. J. Dasher, "Semantics and Kirchhoff's Current Law," Proc. IRE, vol. 47, pp. 1158-1159; June, 1959.
6. S. Okada, Y. Moriwaki, and K. P. Young, "Realization of Boolean Polynomials Based on Incidence Matrices," 1959 Proc. of the Eastern Joint Computer Conference, pp. 120-132; December, 1959.
7. S. Okada and K. P. Young, "Ambit Realization of Cut-Set Matrices into Graphs," report on research sponsored by Air Force Cambridge Research Center, Contract AF-19(604)-6620, performed at Microwave Research Polytechnic Inst. of Brooklyn, New York.
8. P. Dienes, The Taylor Series, Dover Publs. Inc., New York, 1957; Chap. VI.
9. W. H. Kim and R. T. Chien, Topological Analysis and Synthesis of Communication Networks, Columbia University Press, New York; 1962.
10. M. B. Reed, "The Seg: A New Class of Subgraphs," IRE Trans. on Circuit Theory, vol. CT-8, pp. 17-22; March, 1961.
11. E. A. Guillemin, "How to Grow Your Own Trees from Given Cut-Set or Tie-Set Matrices," IRE Trans. on Circuit Theory, vol. CT-6, special supplement, pp. 110-126; May, 1959.
12. A. Shimbel, "Structure in Communication Nets," Proc. of the Symposium on Information Networks, New York, pp. 199-203; April, 1954.
13. N. Jacobson, Lectures in Abstract Algebra, vol. I, Van Nostrand Co., Inc., Princeton, N. J., 1962; p. 20.

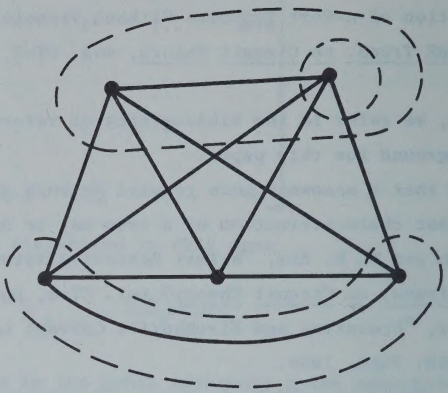


Fig. 1 A Graph and Some Ambits

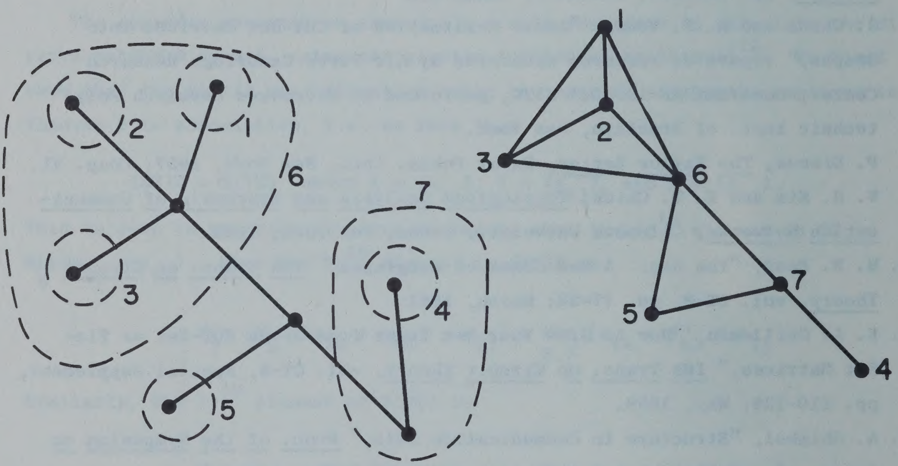
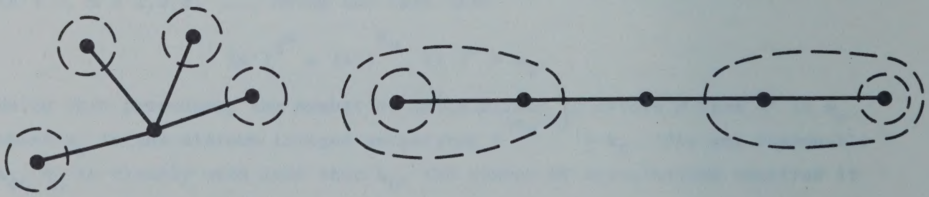


Fig. 2 (a) Basic Ambit Set, (b) Ambit Adjacency Graph



(a) Star Tree

(b) Chain Tree

Fig. 3 Basic Ambit Sets with Extreme Dispersions

FUNDAMENTAL CONNECTION MATRICES*

ANTHONY J. PENNINGTON
Department of Electrical Engineering
The University of Michigan
Ann Arbor, Michigan

Summary

Interest in the applications of linear graph theory to electrical networks has increased rapidly in recent years. The synthesis of transformerless n-port networks has been a particularly important motivating problem. The fast pace of recent work in this area has tended to obscure an important early contribution by Gabriel Kron (Ref. 1), the "orthogonal" method of network analysis. The purpose of this paper is to reinterpret Kron's method in terms of the modern concepts of "fundamental circuits and cut-sets," and to indicate the range of applications of the associated non-singular matrix representations of network topological structures.[1]

Fundamental Connection Matrices

The topological structure of a network of two-terminal elements can be represented by means of an oriented linear graph. Figure 1 is an example of such a graph.

We now define some special terms:

t-graph. A t-graph is an oriented linear graph with a distinguished (specified) tree.

Open path of a t-graph. An open path of a t-graph is any edge of its distinguished tree, i.e., any branch.

---

* This paper is derived from Chapter II of a dissertation submitted by the author in partial fulfillment of the requirements for the degree of Doctor of Philosophy in The University of Michigan, August, 1963, and is based on research supported by the U. S. Air Force, Wright Air Development Command, under Contract No. AF 33(616)-7630, and by the U. S. Army Electronics Command, under Contract No. DA-36-039 sc-89227.

[1] After completion of this paper the author discovered that E. A. Guillemin has carried out essentially the same task from a somewhat different point of view in a recently published book (Ref. 7).

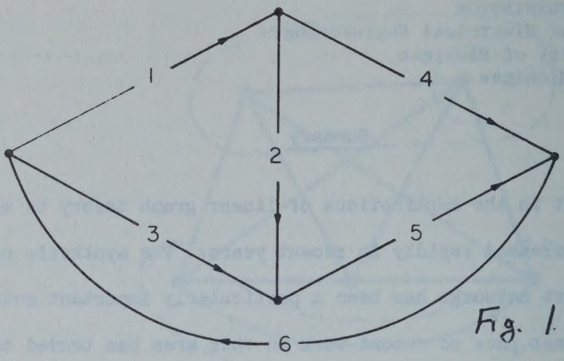


Figure 1. Oriented linear graph (Wheatstone Bridge).

Closed path of a t-graph. A closed path of a t-graph is any edge of the cotree associated with the distinguished tree, i.e., any link, plus the unique set of branches which connect the terminal vertices of the link.

A closed path of a t-graph is identical with the well-known fundamental circuit (Ref. 2). For clarity we shall restrict attention to connected graphs without self loops. Any such graph with e edges and v vertices has

$$b = v - 1 \text{ branches, and}$$
$$\ell = e - b \text{ links.}$$

If Fig. 1 is converted to a t-graph by distinguishing (specifying) the tree formed by edges 4, 5, and 6, then these edges are the open paths and the groups of edges (1, 4, 6), (2, 4, 5), and (3, 5, 6) are the closed paths.

We now define the

Fundamental connection matrix of a t-graph, C_{ee}. The fundamental connection matrix of a t-graph, C_{ee}, is an e x e matrix whose entries, c_{ij} (in row i and column j), are given by the rule:

$$c_{ij} = \begin{cases} 0 & \text{if edge } i \text{ is not a member of path } j; \\ 1 & \text{if edge } i \text{ is a member of path } j\text{, and if traversal of the path in the positive direction of edge } j \text{ results in positive traversal of edge } i. \\ -1 & \text{if edge } i \text{ is a member of path } j\text{, and if traversal of the path in the positive direction of edge } j \text{ results in negative traversal of edge } i. \end{cases}$$

In order to write down C_{ee} it is necessary to adopt some scheme for associating the rows with edges and columns with paths. It is always possible to do this in such a way that the first ℓ rows correspond to the links and the first ℓ columns correspond to the closed paths defined by the links in the same order. Likewise, the last b rows can be made to correspond to the branches, and the last b columns to the open paths defined by the branches in the same order.

A fundamental connection matrix as defined here is a special case of Kron's "connection matrix," which is similar, except that its "paths" need not be defined with respect to a tree. The restriction made here yields a particularly simple matrix representation, as we shall see.

Under the row-column ordering described above C_{ee} always has the form

$$C_{ee} = \begin{bmatrix} U_{\ell\ell} & 0_{\ell b} \\ C_{b\ell} & U_{bb} \end{bmatrix}, \tag{1}$$

where $U_{\ell\ell}$ and U_{bb} are unit matrices, $0_{\ell b}$ is a zero matrix, and $C_{b\ell}$ is a matrix of 0's, 1's, and -1's. The upper-left submatrix is a unit matrix because each link is a member, with positive orientation, of the closed path which it defines, and is a member of no other closed path; the lower-right submatrix is a unit matrix because each branch is a positively oriented member of the open path which it defines, and is a member of no other open path; the upper-right submatrix is a zero matrix because no link is a member of an open path; the lower-left submatrix entries are 0, 1, or -1 depending on the presence and orientation of each branch in the closed paths. Since all submatrices except $C_{b\ell}$ would be the same for any e edge, v vertex t-graph, the connection information in C_{ee} is carried, essentially, by $C_{b\ell}$. The first ℓ columns of C_{ee} constitute the transpose of the usual fundamental circuit matrix.

The fundamental connection matrix for the t-graph obtained by specifying the tree composed of edges 4, 5, and 6 in Fig. 1 is

$$C_{ee} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} . \tag{2}$$

Inverse Fundamental Connection Matrices

The most interesting and useful property of C_{ee} is that it is nonsingular. This fact can be established immediately by noting that C_{ee} is triangular with nonzero diagonal entries, so that the determinant, being the product of the diagonal entries, must be nonzero. Furthermore, the inverse matrix is simply

$$C_{ee}^{-1} = \begin{bmatrix} U_{\ell\ell} & 0_{\ell b} \\ -C_{b\ell} & U_{bb} \end{bmatrix} , \tag{3}$$

as can be checked by multiplying both sides by C_{ee} from the right or left using the usual rule for multiplying partitioned matrices (Ref. 3).

It is possible to interpret this inverse matrix in terms of the associated t-graph by defining the following terms:

Open copath of a t-graph. An open copath of a t-graph is any branch plus the unique set of links which separate the terminal vertices of the branch.

An open copath is identical with a fundamental cut-set (Ref. 2).

Closed copath of a t-graph. A closed copath of a t-graph is any link.

The entries of C_{ee}^{-1} give the membership and orientation of each edge (indexed by the columns) in the open and closed copaths (indexed by the rows). Positive orientation means that the member edge points in the same direction, with respect to the terminal vertices of the edge which defines the copath, as the defining edge itself. The upper-left unit matrix expresses the fact that each link is a positively oriented member of the closed copath which it defines and is a member of no other closed copath; the lower-right submatrix is a unit matrix because each branch is a positively oriented member of the open copath which it defines, and is a member of no other open copath; the upper-right submatrix is a zero matrix because no branch is a member of a closed copath; the lower left submatrix entries are 0, 1, or -1 depending on the membership and orientation of the links in the open copaths. This last submatrix is equal to $-C_{b\ell}$ because every branch which was a positively oriented member of a closed path must define an open copath which includes the closed path defining link in the negative sense.

The last b rows of C_{ee}^{-1} correspond to the usual fundamental cut-set matrix.

This discussion shows that the fundamental connection matrix, C_{ee}, and its inverse, C_{ee}^{-1}, are sufficient to describe the topological characteristics of a network in terms of the dual concepts of fundamental circuits and cut-sets. With this basic tool available circuit analysis on both an impedance and admittance basis becomes straightforward. The word "orthogonal" applied to such analysis by Kron refers to the orthogonality of the rows and columns of the inverse matrices.

Analysis of e-Port Networks

We shall now consider the analysis of a special class of ported networks whose ports are e member sets of terminal pairs, or "e-ports," associated with a t-graph of a network. Later, attention will be focussed on special subsets of these ports, i.e., the customary sets of "n-ports."

Open path ports of a t-graph. The open path ports of a t-graph are the terminal pairs created by connecting to the terminal vertices of its branches.

Closed path ports of a t-graph. The closed path ports of a t-graph are the terminal pairs created by cutting its links.

Open and closed path ports as defined here are identical, respectively, to "soldering iron entry" and "pliers entry" ports (Guillemin, Ref. 4) or to "described" and "inscribed" ports (Cederbaum, Ref. 5), and the new designations serve merely to emphasize the association of each port with open or closed path and with a unique network element.

The impedance and admittance characteristics of a network at the complete set of open and closed path ports can be obtained by simple applications of Ohm's and Kirchhoff's laws expressed in matrix form.

The operational impedance version of Ohm's law for a set of linear inductance, resistance, and elastance elements can be written

$$V_{e1} = Z^{e}_{ee} J_{e1} \tag{4}$$

where V_{e1} is the e x 1 matrix of element potentials, Z^{e}_{ee} is the e x e matrix of element impedances (with any coupling represented by nonzero off diagonal entries), and J_{e1} is the e x 1 matrix of element currents. The positive current, and positive to negative potential rise directions are assigned to agree with the arbitrary element orientations. Kron has treated this equation as a representation of a "primitive" or "disconnected" system. A more satisfactory point of view results from simply recognizing that (4) is a relation which holds among the potentials across and currents through the elements independently of their mode of connection.

The Kirchhoff potential law equation is

$$E_{e1} = (C_{ee})^{t} V_{e1} , \tag{5}$$

where E_{e1} is the e x 1 matrix of voltages across the open and closed ports. The Kirchhoff current law equation is

$$I_{el} = C_{ee}^{-1} J_{el} , \tag{6}$$

where I_{el} is the e x 1 matrix of currents into the ports. The positive current and negative to positive potential rise directions are chosen as the positive orientations of the port-defining elements.

Equations 5 and 6 can be shown to be correct expressions of the Kirchhoff laws by examining the partitioned forms:

$$\begin{bmatrix} E_{\ell 1} \\ E_{b1} \end{bmatrix} = \begin{bmatrix} U_{\ell\ell} & (C_{b\ell})^t \\ 0_{b\ell} & U_{bb} \end{bmatrix} \begin{bmatrix} V_{\ell 1} \\ V_{b1} \end{bmatrix} , \tag{7}$$

and

$$\begin{bmatrix} I_{\ell 1} \\ I_{b1} \end{bmatrix} = \begin{bmatrix} U_{\ell\ell} & 0_{\ell b} \\ -C_{b\ell} & U_{bb} \end{bmatrix} \begin{bmatrix} J_{\ell 1} \\ J_{b1} \end{bmatrix} . \tag{8}$$

The first row of (7) gives the voltage balance around the closed paths or fundamental circuits, while the second expresses the fact that the open path port and branch voltages are identical. The first row of (8) states that the closed path port and link currents are identical, while the second row gives the current balance across the open copaths or fundamental cut-sets.

Multiplying both sides of Eq. 6 by C_{ee} and incorporating the result, along with Eq. 5, in Eq. 4 gives

$$E_{el} = (C_{ee})^t Z_{ee}^e C_{ee} I_{el}$$

$$= Z_{ee}^p I_{el} , \tag{9}$$

where

$$Z_{ee}^p = (C_{ee})^t Z_{ee}^e C_{ee} \tag{10}$$

is the matrix of open and closed path port impedance characteristics.

Multiplication of both sides of Eq. 5 by $[(C_{ee})^t]^{-1}$ yields an equation which, along with Eq. 6, permits us to rewrite the element admittance equation,

$$J_{el} = Y_{ee}^e V_{el} \tag{11}$$

as

$$I_{el} = C_{ee}^{-1} Y_{ee}^e [(C_{ee})^t]^{-1} V_{el}$$

$$= Y_{ee}^p V_{el} , \tag{12}$$

where

$$Y^{p}_{ee} = C^{-1}_{ee}Y^{e}_{ee}[(C_{ee})^{t}]^{-1} \tag{13}$$

is the matrix of port admittance characteristics.

Equation 13 can also be obtained directly by inverting Eq. 10, provided that the inverse of the element impedance matrix, $(Z^{e}_{ee})^{-1} = Y^{e}_{ee}$, exists.

Power Invariance

In a recent publication (Ref. 6) Kron has raised the problem of interpreting the principle of "power invariance" under transformations from disconnected to connected networks. A correct interpretation is readily derivable from the Kirchhoff laws as presented in Eqs. 5 and 6. The complex input power at the e-ports is,

$$\begin{aligned}(E^{*}_{el})^{t}I_{el} &= \{[(C_{ee})^{t}V_{el}]^{*}\}^{t}C^{-1}_{ee}J_{el} \\ &= (V^{*}_{el})^{t}J_{el} \quad . \end{aligned} \tag{14}$$

The expression on the right is simply the complex power absorbed by the elements, so that we have the obvious conservation of energy statement for connected networks without the troublesome problem of physically interpreting what happens during the process of "connection."

The fact that C_{ee} is nonsingular also makes it easy to obtain either of the Kirchhoff laws from the other one plus the power invariance statement, so that any two of the three principles serve as an adequate basis for network analysis.

Analysis of n-Port Networks

We shall now treat the conventional problem of obtaining the immittance characteristics of a network at some subset of its e-ports, i.e., at a set of "n-ports." In order to do this it is necessary to apply proper constraints at the remaining e-ports; in particular, these ports must be "passive" (exchange zero power, instantaneously, with the network) and the topological structure of the original network must be unchanged. These requirements can be met by requiring that the potentials across any remaining closed path ports and the currents into any remaining open path ports be set equal to zero. These operations correspond to short-circuiting the remaining closed path ports and open-circuiting the remaining open path ports. As an example, the impedance characteristics at an n member set of the closed ports can be obtained by applying the constraints,

$$E_{ml} = 0_{ml}\,, \tag{15}$$

and

$$I_{bl} = 0_{bl}\,, \tag{16}$$

where E_{ml} is the matrix of remaining closed path port potentials and I_{bl} is the matrix of open path port currents. Under these conditions the partitioned form of Eq. 10 becomes

$$\begin{bmatrix} E_{nl} \\ 0_{ml} \\ E_{bl} \end{bmatrix} = \begin{bmatrix} Z^p_{nn} & Z^p_{nm} & Z^p_{nb} \\ Z^p_{mn} & Z^p_{mm} & Z^p_{mb} \\ Z^p_{bn} & Z^p_{bm} & Z^p_{bb} \end{bmatrix} \begin{bmatrix} I_{nl} \\ I_{ml} \\ 0_{bl} \end{bmatrix} \tag{17}$$

Expanding and eliminating I_{ml} between the first two rows of the result gives

$$E_{nl} = [Z^p_{nn} - Z^p_{nm}(Z^p_{mm})^{-1}Z^p_{mn}]I_{nl} \tag{18}$$

so that the matrix in brackets on the right-hand side is the desired n-port impedance matrix with passivity and structure preserving constraints applied at the remaining e-ports. A similar development can be used to obtain the admittance characteristics at any desired n member subset of the open path ports.

Implications for General Network Studies

This paper has presented and interpreted a special version of Kron's nonsingular "connection matrices," and indicated their application to the analysis of ported networks. The importance of nonsingular topological matrix representations, however, extends far beyond the simplification of network analysis exhibited here. The existence of an inverse of C_{ee} opens up possibilities for generating entirely new results in network theory. In particular, it is now possible to begin determining conditions on the separate topological and "electromagnetic" structures of connected networks for various performance characteristics. A forthcoming publication by the author will present comprehensive results of this type for general n-port equivalent networks.

Looking further ahead, it seems clear that fundamental progress on the general n-port synthesis problem will require a deeper understanding of the roles played by the separate structures than is now available.

Acknowledgment

The author thanks his doctoral committee chairman, Professor E. L. McMahon, for his valuable guidance throughout this research effort.

References

1. Gabriel Kron, "Tensor Analysis of Networks," John Wiley and Sons, Inc., New York, N. Y., pp. 407-435; 1939.
2. Sundaram Seshu and Myril B. Reed, "Linear Graphs and Electrical Networks," Addison-Wesley Publishing Co., Inc., Reading, Massachusetts, pp. 24-33; 1961.
3. A. M. Tropper, "Matrix Theory for Electrical Engineers," Addison-Wesley Publishing Co., Inc., Reading, Massachusetts, pp. 37-39; 1962.
4. Ernst A. Guillemin, "Synthesis of Passive Networks," John Wiley and Sons, Inc., New York, N. Y., p. 172; 1957.
5. Israel Cederbaum, "Conditions for the Impedance and Admittance Matrices of N-Ports Without Ideal Transformers," Proceedings of the Institution of Electrical Engineers (London), Monograph 276R, January, 1958.
6. Gabriel Kron, "What is an Electric Network?" The Matrix and Tensor Quarterly (Correspondence), Vol. 13, No. 3, p. 21; March, 1963.
7. Ernst A. Guillemin, "Theory of Linear Physical Systems," John Wiley and Sons, Inc., New York, N. Y., pp. 12-20; 1963.

EVALUATION OF NETWORK DETERMINANTS VIA THE KEY SUBGRAPH*

S. D. BEDROSIAN
Institute for Cooperative Research
University of Pennsylvania
Philadelphia, Pennsylvania

Abstract - In linear network theory there is frequency need for evaluation of network determinants. A simple technique of evaluation is presented based on the key subgraph of the network. It systematically yields a new set of partially factored forms which yield explicit results. By establishing generic forms for the key subgraph, general formulas are obtained for the total number of (trees) terms in the expansion. These are further related by recursion formulas.

1. INTRODUCTION

In linear network theory there is frequent need for evaluation of network determinants. A simple technique of evaluation is presented based on the key subgraph of the network. The concept of the key subgraph was introduced in connection with network solvability investigations.[1] The key subgraph of a linear graph representation of a network is defined as the set of branches incident on any internal or concealed nodes. This key subgraph takes into account the disposition as well as the number of internal nodes. In addition to providing a means for reducing the number of equations representing short-circuit admittance measurements to a key subset, the key subgraph provides a means for categorizing networks based on the number of internal nodes, I, and the concealed branches, B_c, between them as parameters. This is presented in terms of generic forms. An additional off-shoot of the key subgraph concept has been the development of compact graphical representations for evaluating the nodal determinant of a network. These partially factored forms are elementary hinged networks which differ from any of those previously given. Furthermore, they are simple graphs whose trees may be written simply by inspection. It systematically yields a new set of partially factored forms which yield explicit results. These are further related by recursion formulas.

II. GENERIC FORMS OF KEY SUBGRAPH

We can enlarge upon our earlier mention of the application of key subgraphs to the problem of categorization of networks.[2] Investigation of this problem using the number of internal nodes and concealed branches as parameters resulted in the pattern for key subgraph forms shown in Table I. Since each entry in the table does not necessarily represent a unique form we circumvent the problem by displaying only generic forms of the key subgraph. This is shown in Figure 1 for the case of two non B_c branch at each node.

---

* This paper is an outgrowth of work done under Contract DA-36-034-507-ORD-3347A between the Institute for Cooperative Research and the Frankford Arsenal, Army Munitions Command.

TABLE I, Initial Categorization of Networks

Nodes I	Number of Concealed Branches B_c
2	1
3	2 3
4	3 4 5 6
5	4 5 6 7 8 9 10
6	5 6 7 8 9 10 11 12 13 14 15
.	.
.	.
.	.
I	(I-1) in integral steps to I(I-1)/2

III. LINEAR CASE

With this basic framework in mind we can now indicate the general pattern of the partially factored forms applicable to the "linear" cases. Brief discussion of some compact topological representations obtained from the key subgraph appeared in a previous paper.[3] Careful study of the partially factored forms for the linear cases reveals that the number of forms obtained is

$$2^{B_c}$$

which can be expressed as the sum of binomial coefficients. Such a tabulation is shown in Table II.

TABLE II, Linear Case of Key Subgraph

	Number of Concealed Branches B_c							Trees
Nodes I	0	1	2	3	4	5	6	$T_{I}^{2\ell}$,
2	1	1						(8)
3	1	2	1					(30)
4	1	3	3	1				(112)
5	1	4	6	4	1			(418)
6	1	5	10	10	5	1		(1560)
7	1	6	15	20	15	6	1	(5822)
.								.
.								.
.								.

The entries in this chart show the number of compact forms having the prescribed number of concealed branches. The numbers in parenthesis in the last column on the right are obtained by summation of the trees for these forms and give the total number of trees for each case of the generic forms illustrated in the first column of Figure 1. In Figure 2 are shown examples of the compact forms for a few specific cases. Note that each of the partially factored forms contains the product of I terms when expanded. This is best seen by examining the case of the compact representations for the determinant of a matrix in tridiagonal form. Figure 3a shows a simple three node generic graph for $\nu = 1$. This graph is redrawn in Figure 3b with an explicit reference node. The nodal admittance matrix corresponding to the branch network of Figure 3 with each branch taken as a unit admittance may be written as follows:

$$\begin{array}{c|c|c|c|} \text{(Node)} & 1 & 2 & 3 \\ \hline 1 & 2 & -1 & \\ \hline 2 & -1 & 3 & -1 \\ \hline 3 & & -1 & 2 \\ \hline \end{array}$$

The determinant representation of this matrix in terms of the partially factored forms is shown in Figure 4. Note that if literal values are assigned to the branches of the graph and the partially factored forms the determinant expansion can be written explicitly. Then the terms in the determinant expansion are:

$$T = abc + xc(a+b) + ya(b+c) + xy(a+b+c). \qquad (1)$$

If we substitute unit value for each of the elements we get the following expression for the total number of terms

$$T = 1 + 2 + 2 + 3 = 8. \qquad (2)$$

The topological representations which pertain to this example are of course related to those for the linear case of $I = 3$ shown in Figure 2.

IV. BASIC LOOP CASE

We next consider the number of compact forms for the basic loop cases of key subgraphs. These cases relate to the second column of generic forms shown in Figure 1. Note in the table which follows that there are

$$2^{Bc} - 1$$

forms required. These are expressed by means of a truncated binomial expansion.

As in the previous table the values in terms of the total number of trees are given in parenthesis in the last column. Some examples of the compact forms for a few basic loop cases are shown in Figure 5.

TABLE III, Basic Loop Case of Key Subgraph

Nodes I	Number of Concealed Branches B_c							Trees
	0	1	2	3	4	5	6	T_I^{2m}
2	1	2						(12)
3	1	3	3					(50)
4	1	4	6	4				(192)
5	1	5	10	10	5			(722)
6	1	6	15	20	15	6		(2,700)
7	1	7	21	35	35	21	7	(10,082)
.	.							
.	.							
.	.							

V. RECURSION FORMULAS

In order to facilitate our discussion of the total number of trees for various cases we introduce subscript and superscript notation along with the symbol T for the total number of trees. The subscript I indicates that the total number of trees relates to the case of a key subgraph with I nodes. The superscripts are in general ν, ℓ or ν, m where the ν refer to the number of non B_c branches at each node of the graph and the ℓ and m refer to the linear or basic loop cases respectively. Careful investigation of the systematic procedure for generating the pertinent partially factored forms suggest that the totals obtained for each row in Table II and Table III must be related to the preceding or subsequent rows. Without detailed discussion of the steps involved we give the recursion formula for the total number of trees for the linear cases of Table II as follows:

$$(\nu + 2)T_I^{\nu,\ell} - T_{I-1}^{\nu,\ell} = T_{I+1}^{\nu,\ell} \qquad \nu=1,2,3\ldots \qquad (3)$$

The form of Equation (3) will suggest to the interested reader reasonable steps for obtaining it. One finds that a similar procedure will produce a recursion formula which will relate the total number of trees between adjacent rows of Table III. This equation follows.

$$(v+2)T_I^{\nu,m} - T_{I-1}^{\nu,m} + 2\nu = T_{I+1}^{\nu,m} \qquad \nu=1.2,3\ \ldots \qquad (4)$$

The special case of equation (2) for $\nu = 1$ was known to Professor R. M. Foster[4] in connection with (I + 1) node triply-connected graphs having 2(I + 1) branches.

It is worth stressing at this point that the trees for the partially factored form relate to the evaluation of the determinant of the network for which the key subgraph is a compact representation. The notation I is retained for nodes although we no longer constrain ourselves to internal nodes. The basic notion of using a tree expansion of a linear graph as a means of evaluation determinants is well known. Observe, however, that the systematic procedure described here has resulted in recursion formulas which have implications on determinant relationships. Using a different basis, Pullen[5] gave a recursion formula for a ladder network which should be related to equation (1) for $\nu = 1$.

What is perhaps even more interesting is the fact that further scrutiny of the systematic procedure for formulating the compact representations as well as the recursion formulas given in Equations (3) and (4) has permitted establishment of an additional interrelationship. Before giving the resulting recursion formula we pause to examine some numerical results obtained for the linear and basic loop cases both for $\nu = 2$. This is shown in Figure 6. Observe that this is equivalent to a closer look at the first two columns of the generic forms given in Figure 1. For completeness, we have also added the degenerate cases for $I = 1$. Offhand one would have no reason to suspect that the numbers representing the total number of trees in one column are related to those of the other. Examination of any two successive rows however reveal that the following formula holds.

$$T_I^{2,\ell} + T_I^{2,m} + T_{I+1}^{2,\ell} = T_{I+1}^{2,m} \qquad (\nu = 2) \qquad (5)$$

Then for the general case we have the recursion formula:

$$T_I^{\nu,\ell} + T_I^{\nu,m} + T_{I+1}^{\nu,\ell} = T_{I+1}^{\nu,m} \qquad \nu = 1,2,3 \ldots \qquad (6)$$

Thus Equation (6) suggests the possibility of some new determinant relations. These results also bear on the problem of network analysis by digital computer. A different approach to that just presented is being pursued at the Research Laboratory of Electronics at M.I.T.[6] It also seems to tie in with some earlier work by Parter[7] as well as that by Chen and Haas.[8]

VI. DISCUSSION

It was pointed out at the outset that the key subgraph concept evolved during the development of a systematic procedure for the formulation of equations expressing the externally observable behavior of a multiterminal network having some concealed nodes. The key subgraph was related to a submatrix of a partitioned indefinite admittance matrix.[1] In particular the entries in the submatrix related to the internal nodes of the network, when considered from a topological point of view, were found to represent a special subgraph of the given network. This key subgraph was shown to yield a new partially factored form representation for the system determinant of the indefinite admittance matrix.[3] In addition to that application we find that the generic forms can be interpreted as relating to the graph of a simple network with its reference or datum node not explicitly shown. In a sense the generic graph is comparable to the conventional node admittance matrix as opposed to the indefinite admittance matrix of a branch network. Viewed in this light the simpler forms are readily recognized. In particular the so-called linear case of our generic forms are related to the familiar ladder networks. Consider the linear case of four internal nodes with three concealed branches using the typical case previously shown in Figure 2 for $\nu = 2$. The matrix for such a graph, with literal branch designations and with all the non B_c branches joined at a common reference node, may be written as

a+b+x	-x		
-x	c+d+x+y	-y	
	-y	e+f+y+z	-z
		-z	g+h+z

Then the determinant Δ may be expressed in the partially factored form:

$$
\begin{aligned}
\Delta = {} & (a+b)(c+d)(e+f)(g+h) \\
& +x(e+f)(g+h)(a+b+c+d) \\
& +y(a+b)(g+h)(c+d+d+f) \\
& +z(a+b)(c+d)(e+f+g+h) \\
& +yz(a+b)(c+d+e+f+g+h) \\
& +xz(a+b+c+d)(e+f+g+h) \\
& +xyz(a+b+c+d+e+f+g+h)
\end{aligned}
$$

The corresponding topological compact forms were shown in Figure 2. In either case the total number of terms (trees) in the determinant expansion is 112.

The next example is also for a linear case of the generic forms. This time we use $I = 4$ as shown in Figure 7. The node admittance matrix corresponding to this network is, for unit admittances;

3	-1		
-1	3	-1	
	-1	3	-1
		-1	3

Observe that the resulting matrix is in tridiagonal form with all the diagonal elements equal. This resulted from the use of $\nu = 2$ for the two outside branches of the network (1-REF and 2-REF), while retaining $\nu = 1$ for the inner branches in Figure 7. Using the compact forms for evaluating the determinant of this matrix, we find the results shown in Figure 8. The total number of terms in the determinant expansion is given by

$$T = 4 + [2(6) + 8] + [2(8) + 9] + 6 = 55 \qquad (8)$$

This second example was included to emphasize the versatility of the partially factored forms. Although the generic forms originally presented are symmetrical the basic technique is not limited in this way.

Another demonstration of the fundamental nature of this technique will be shown. Given a graph in the loop case by means of our compact forms we can "see" how this graph can be made to revert to a graph in the linear category. Consider the graph and its expanded forms shown in Figure 9. Let the w-branch in each of the compact forms take on the value of zero admittance. Then the forms containing w must vanish since the corresponding product of admittances becomes zero. Thus for this case of four nodes we can compare the results given previously in Tables II and III as shown in Table IV below.

TABLE IV, Comparison of Four Node Key Subgraphs

Case	Number of B_c Branches 0	1	2	3	Total Forms	Total Trees
Loop	1	4	6	4	15	192
Linear	1	3	3	1	8	112

To simplify the comparison the pertinent forms which are deleted in this conversion are indicated in Figure 9.

We summarize a systematic procedure for expanding the key subgraph into the compact topological representations which express the determinant of the network in partially factored form. In other words this is an equivalent explicit statement of all the trees of the network. Tables I and II list, for a given number of nodes in the key subgraph, the breakdown for the number of compact forms which include a specified number of concealed branches. To obtain the compact forms having

$B_c = 0$

Join the non B_c branches at each node so that they are in parallel. Then collapse all concealed branches so that all the nodes are merged (coalesced) into a single node.

$B_c = 1$

All branches incident at a given node are joined at the reference end, except that for the selected B_c branch, the set of all non B_c branches at both of its nodes are joined so that they are in parallel. Collapse all remaining concealed branches. This process is repeated until each B_c branch has been used to construct a compact form.

$B_c = 2$

The same basic procedures are followed as above for $B_c = 1$, except that the B_c branches are now selected two at a time Furthermore there are two possible cases. The set of incident branches to be considered will either include four nodes, if the selected B_c branches are not contiguous, or only three nodes, if the selected branches are incident to a common node.

[for $B_c \geq 2$ the B_c branches selected for a compact form must not form a closed loop e.g., 2-cycle, 3-cycle, etc.]

.

.

.

$B_c = I-1$

The procedure terminates when all of the B_c branches are taken I-1 at a time. However, the I-1 such branches selected must be a tree of the subset of all B_c branches of the key subgraph. In this case the set of all the incident non B_c branches are joined in parallel at a common node and each of the I-1 chosen B_c branches is included in the final hinged compact form.

Two simple checks are available to detect errors.

1. The number of forms of a given class must be as tabulated.
2. The total number of hinged branches will be the same for all forms and will be equal to the number of nodes in the key subgraph.

Because of the systematic nature of the development into partially factored forms it has also been possible to obtain general formulas for the total number of terms in the determinant expansion. These are illustrated in figures 10 and 11 for the linear and basic loop cases. For convenience in writing the formulas n was used instead of ν for the variable.

VII. CONCLUSIONS

The techniques described for evaluation of network determinants offer some new insights into this long standing problem. The symmetrical generic forms were selected as being suitable to illustrate a basic pattern which can serve as the basis for specific investigations. In any case they should not be misinterpreted as implying a limitation on the versatility of the partially factored forms. The examples were chosen with this in mind. Another advantage of the techniques presented is that one can visualize the effect of changes in the individual compact forms used to represent the determinant expansion.

The systematic procedure described yields explicit results for specific cases and lends itself to being programmed on a digital computer. General formulas were given for the total number of terms (trees) in the expansion for two categories of generic key subgraphs Simple recursion formulas were also given for these categories. Further extension of this work should yield additional results with even broader implications.

VIII. ACKNOWLEDGEMENT

Support of expenses in connection with this paper came in part from the Moore School of Electrical Engineering of the University of Pennsylvania.

IX. REFERENCES

1. S. D. Bedrosian and R. S. Berkowitz, "Solution Procedure for Single-Element Kind Networks", IRE International Record, Part 2, pp. 16-24, 1962.
2. C. Beckman, et al, "Study of Piece-Part Fault Isolation by Computer Logic," Status Report, Institute for Cooperative Research, University of Pennsylvania, Contract No. DA-36-034-507-ORD-3347A, June 1963.
3. S. D. Bedrosian, "Simplified Explicit Solution of Networks with Two Internal Nodes", Conference Paper CP 63-968, IEEE Summer General Meeting and Nuclear Radiation Effects Conference, Toronto, Canada, June 1963. Also to appear IEEE Transactions.
4. R. M. Foster, Private communication, May 1963.
5. K. A. Pullen, Jr. "On the Properties of Ladder Networks", Ballistic Research Laboratories, Report 1102, Aberdeen Proving Ground, Maryland, April 1960.
6. D. U Wilde, "Network Analysis by Digital Computer", Quarterly Progress Report No. 70, Research Laboratory of Electronics, M.I.T., July 15, 1963, pp. 307-308.
7. S. Parter, "Use of Linear Graphs in Gauss Elimination", SIAM Review, Vol. 3, No. 2, April 1961, pp. 119-130.
8. C. F. Chen and I. J. Haas, "An Electrical Interpretation of the Algorithm of Gauss", IRE Transactions on Circuit Theory, Vol. CT-9, No. 3, September 1962, pp. 398-399.

Figure I. GENERIC FORMS OF KEY SUBGRAPH

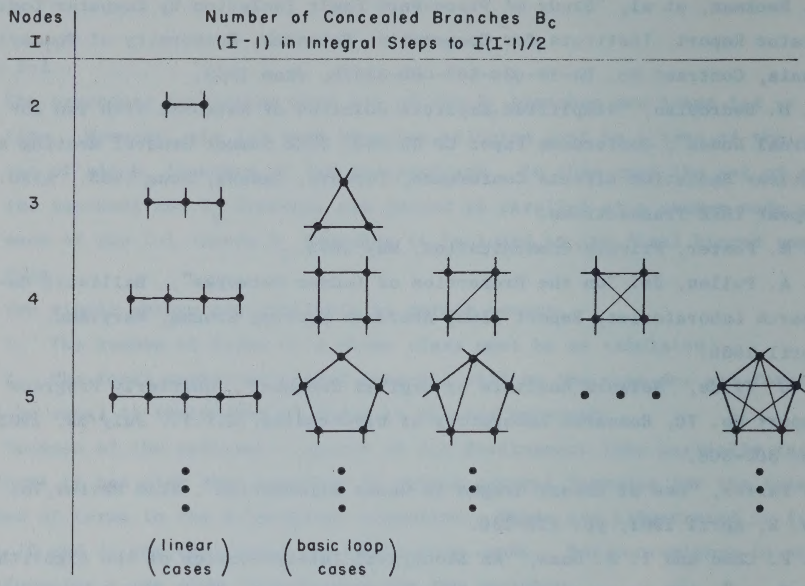
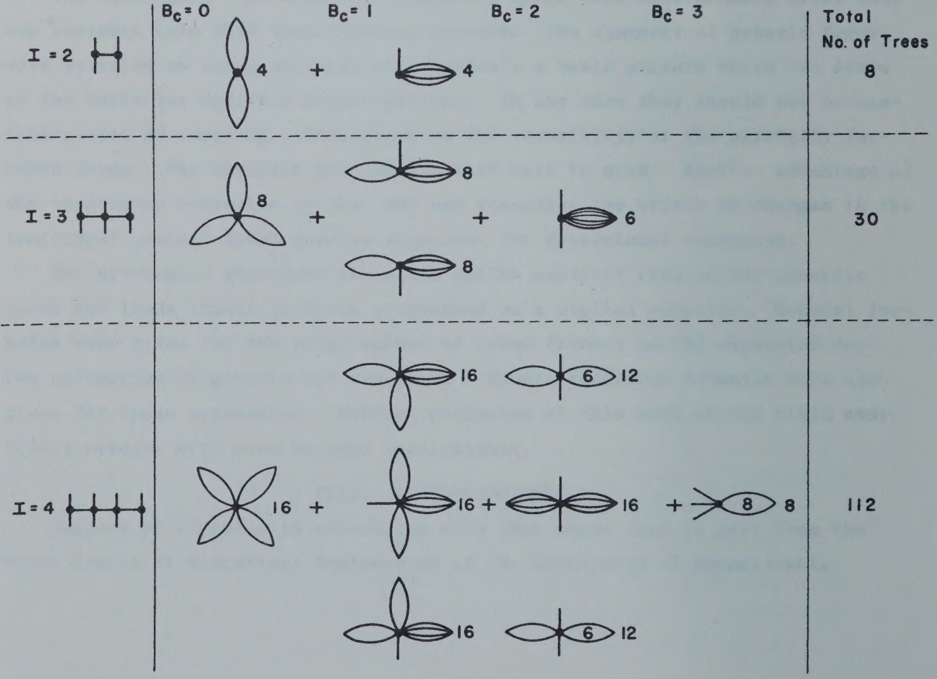

Nodes
I
Number of Concealed Branches B_c
(I - 1) in Integral Steps to I(I-1)/2
2
3
4
5
linear cases
basic loop cases

Figure 2. COMPACT REPRESENTATION OF THE KEY SUBGRAPH IN THE LINEAR CASE


B_c = 0
B_c = 1
B_c = 2
B_c = 3
Total
No. of Trees
I = 2
4
+
4
8
I = 3
8
+
8
+
6
8
30
16
6
12
I = 4
16
+
16 +
16
+
8
8
112
16
6
12

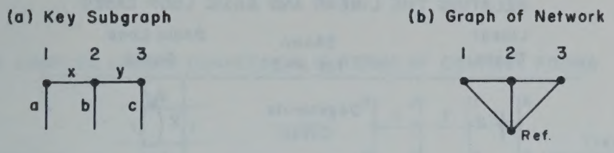
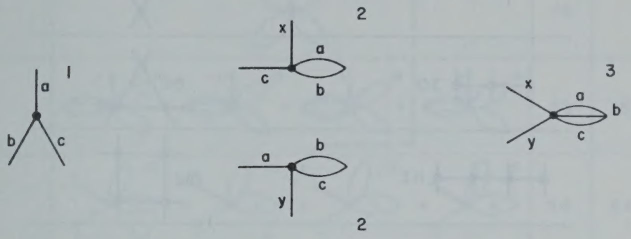

(a) Key Subgraph
1
2
3
x
y
a
b
c
(b) Graph of Network
1
2
3
Ref.

Figure 3. SAMPLE NETWORK


1
a
b
c
2
x
a
c
b
b
a
c
y
2
3
x
a
b
y
c

$$T = abc + xc(a+b) + ya(b+c) + xy(a+b+c)$$

for unit admittances $T = 1 + 2(2) + 3 = 8$

Figure 4. COMPACT FORMS FOR GRAPH OF Figure 3

Figure 5. COMPACT REPRESENTATION OF THE KEY SUBGRAPH WITH BASIC LOOP CASE

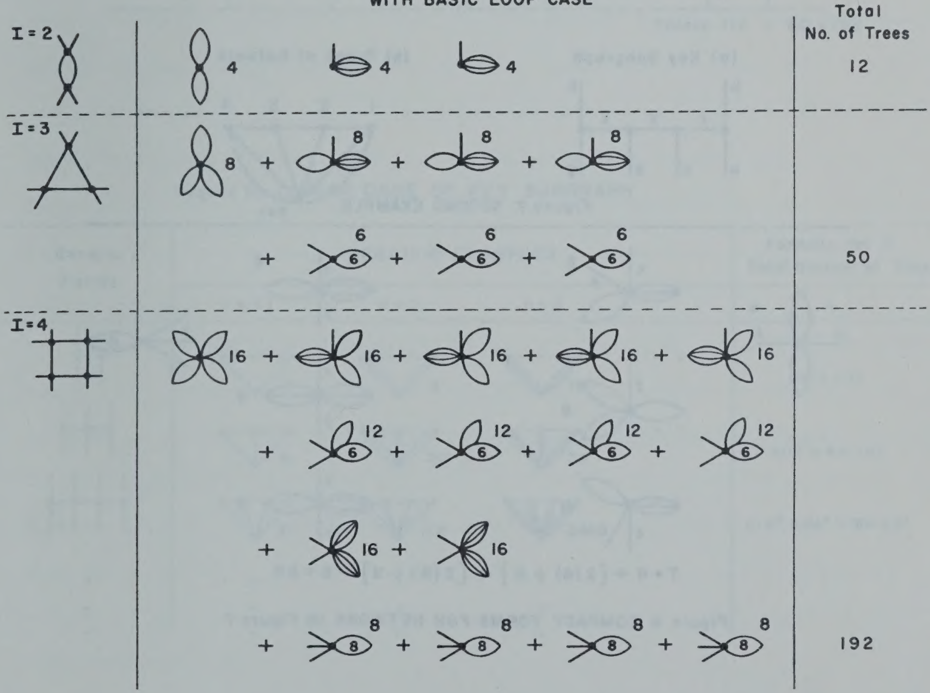
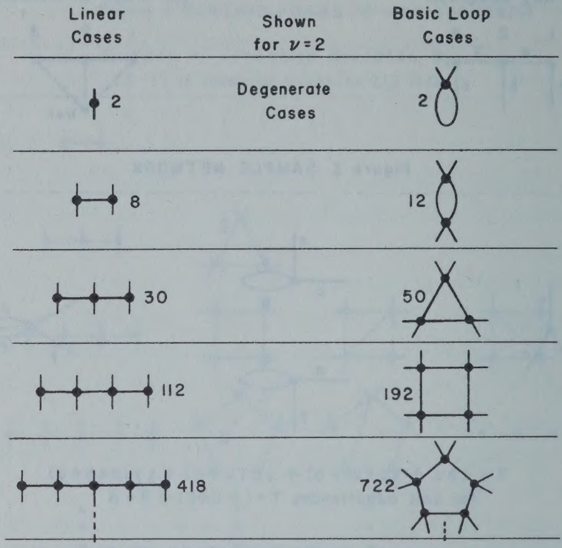
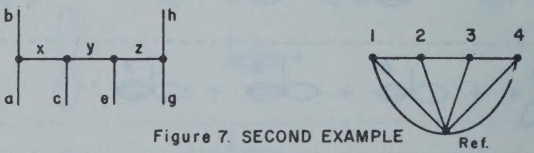

Total No. of Trees
I=2
4
4
4
12
I=3
8
+
8
+
8
+
8
+
6
6
+
6
6
+
6
6
50
I=4
16
+
16
+
16
+
16
+
16
+
12
6
+
12
6
+
12
6
+
12
6
+
16
+
16
+
8
8
+
8
8
+
8
8
+
8
8
192

Figure 6. RECURSION FORMULAS FOR KEY SUBGRAPHS RELATING THE LINEAR AND BASIC LOOP CASES

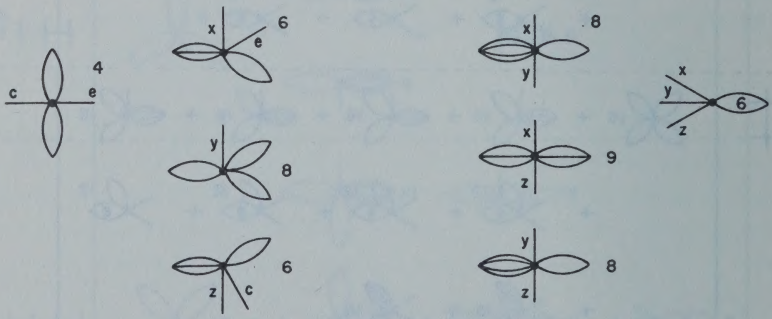

Linear Cases
Shown for ν = 2
Basic Loop Cases
2
Degenerate Cases
2
8
12
30
50
112
192
418
722

$$T_I^{\nu,m} + T_I^{\nu,\ell} + T_{I+1}^{\nu,\ell} = T_{I+1}^{\nu,m} \qquad \nu = 1, 2, 3 \ldots$$


b
h
x
y
z
a
c
e
g
1
2
3
4
Ref.

(a) Key Subgraph (b) Graph of Network

Figure 7. SECOND EXAMPLE


c
e
4
x
e
6
y
8
z
c
6
x
y
8
x
z
9
y
z
8
x
y
z
6

$$T = 4 + [2(6) + 8] + [2(8) + 9] + 6 = 55$$

Figure 8. COMPACT FORMS FOR NETWORK IN Figure 7

Figure 9. LOOP TO LINEAR CONVERSION IN TERMS OF COMPACT FORMS

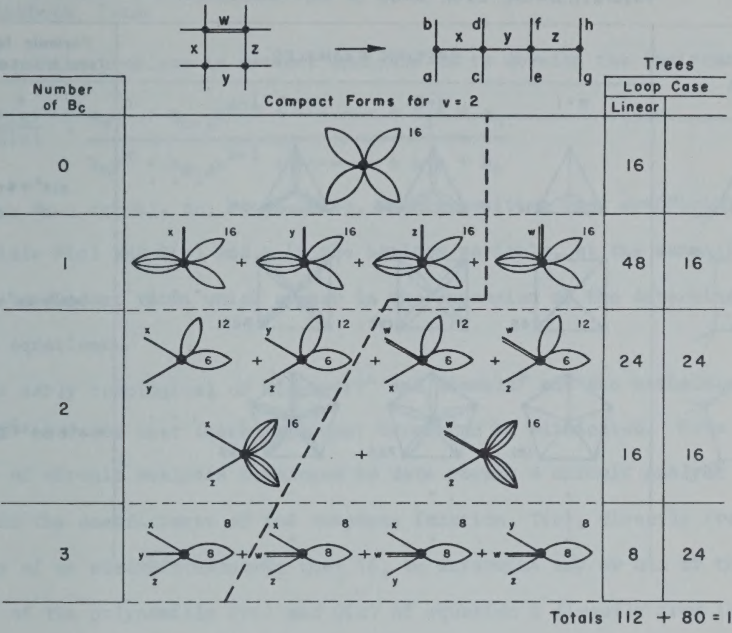

w
x
z
y
b
d
f
h
x
y
z
a
c
e
g
Number of B_c
Compact Forms for v = 2
Trees
Loop Case
Linear
0
16
16
1
x
y
z
w
16
48
16
2
12
6
24
24
16
16
16
3
8
8
24
Totals 112 + 80 = 192

Figure 10. LINEAR CASE OF KEY SUBGRAPH

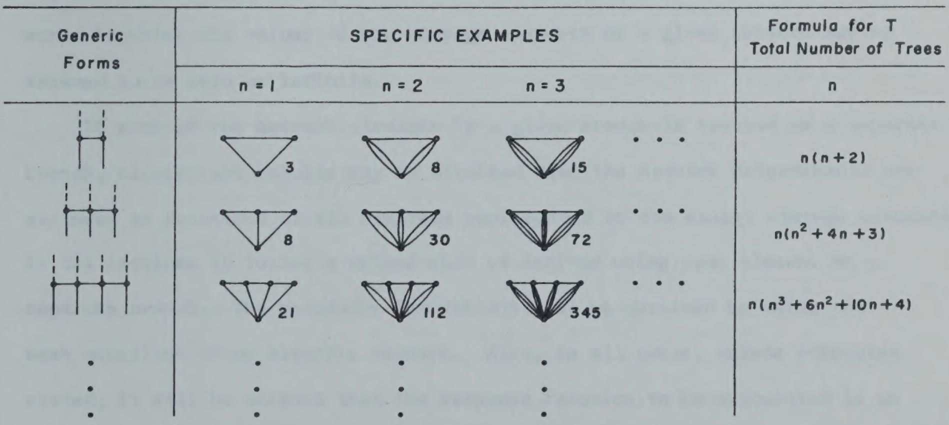
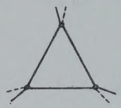
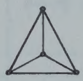
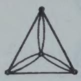
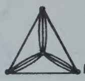
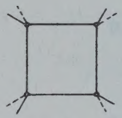
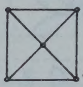
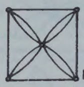
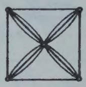
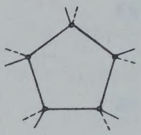
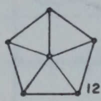
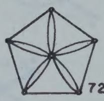
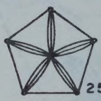

Generic Forms
SPECIFIC EXAMPLES
n = 1
n = 2
n = 3
Formula for T Total Number of Trees
n
3
8
15
n(n+2)
8
30
72
$n(n^2+4n+3)$
21
112
345
$n(n^3+6n^2+10n+4)$

Figure 11. BASIC LOOP CASE OF KEY SUBGRAPH

Generic Forms	SPECIFIC EXAMPLES				Formula for T Total Number of Trees
	n = 1	n = 2	n = 3		n
	 16	 50	 108	. . .	$n(n^2+6n+9)$
	 45	 192	 525	. . .	$n(n^3+8n^2+20n+16)$
	 121	 722	 2523	. . .	$n(n^4+10n^3+35n^2+50n+25)$
⋮	⋮	⋮	⋮		

A METHOD OF NETWORK ANALYSIS USING RESIDUAL NETWORKS

R. H. SEACAT
Texas Technological College
Lubbock, Texas

One of the problems in network analysis is to develop the immittance function

$$T(s) = \frac{P(s)}{Q(s)} = \frac{a_n s^n + a_{n-1} s^{n-1} + \text{-------} + a_1 s + a_o}{b_m s^m + b_{m-1} s^{m-1} + \text{-------} + b_1 s + b_o} \qquad (1)$$

where a_n, a_{n-1} ----a_1, a_o, b_m, ---- b_1, b_o are positive real coefficients of the polynomials P(s) and Q(s) and s is the Laplace variable, at the same time eliminating the redundant terms which appear in the expansion of the determinant of the network equations.

The early topological of Kirchhoff[4] and Maxwell[5] and the methodology of Percival[6] indicate that these redundant terms can be eliminated. None of the methods of circuit analysis developed to date permit a circuit analyst to ascertain the coefficients of the response function, T(s), directly from the topology of an electric network; that is, to determine any or all of the a's and b's of the polynomials P(s) and Q(s) of equation 1 directly from the structure of the network. As a consequence, it is the intent of this paper to describe a method by which the a's and b's in equation 1 can be determined by direct inspection of the network topology. The techniques to be developed in the following sections will be dependent upon the concept of residual networks: those networks in which the values of the storage elements of a given network may be assumed to be zero or infinite.

If each of the network elements in a given branch is treated as a separate branch, significant results may be obtained when the network determinants are expanded as functions of all branches represented by the energy storage elements. In the sections to follow a method will be derived using each element as a separate branch. The necessary derivations will be obtained by using the mesh equations of an electric network. Also, in all cases, unless otherwise stated, it will be assumed that the response function to be calculated is an admittance function.

*This paper is based on part of a dissertation for the degree of Ph.D. at Agricultural and Mechanical College of Texas, College Station, Texas, 1963.

THE NETWORK FUNCTION

There is a relationship between the determinants of the numerator and denominator of a network admittance function which will play an important role in the development of the method of residual networks. Suppose that the principal determinant of a network with m branches is expanded as a function of any three of the impedance elements, Z_1, Z_2, and Z_3, the equation for the expansion will be:

$$\Delta = Z_1Z_2Z_3\Delta_{123} + Z_1Z_2\Delta_{123'} + Z_2Z_3\Delta_{1'23} + Z_1Z_3\Delta_{12'3} + Z_1\Delta_{12'3'} + Z_2\Delta_{1'23'} + Z_3\Delta_{1'2'3} + \Delta_{1'2'3'} \qquad (2)$$

where Δ is the principal determinant and the minors on the right side are derived from the principal determinant in the usual manner. For example, the determinant, $\Delta_{12'3'}$, is obtained from the principal determinant by collapsing the row and column containing Z_1, and setting Z_2 and Z_3 equal to zero. If a single source, E_1, exists in the ith mesh, the equation for Δ_{ik} will be:

$$\Delta_{ik} = Z_1Z_2Z_3\Delta_{ik123} + Z_1Z_2\Delta_{ik123'} + Z_2Z_3\Delta_{ik1'23} + Z_1Z_3\Delta_{ik12'3} + Z_1\Delta_{ik12'3'} + Z_2\Delta_{ik1'23'} + Z_3\Delta_{ik1'2'3} + \Delta_{ik1'2'3'} \qquad (3)$$

where the subscript ik for all Δ's designates the fact that the ith row and kth column of the principal determinant are collapsed. An inspection of equations 2 and 3 yields the fact that the coefficient of a given Z-product in the numerator is a co-factor of the corresponding Z-product in the corresponding Z-product in the denominator. For example, the minors in equations 2 and 3 which are multiplied by the product of Z_2Z_3 are $\Delta_{1'23}$ and $\Delta_{ik1'23}$. $\Delta_{ik1'23}$ may be interpreted as being the co-factor of the determinant $\Delta_{1'23}$. This relationship will be referred to in the calculation of the numerator of admittance functions containing storage elements.

CALCULATION OF THE DENOMINATOR

Suppose that it is desired to expand the network determinant as a function of some inductive element, L_1, the equation for the principal determinant, Δ, will be:

$$\Delta = sL_1\Delta_1 + \Delta_{1'}. \qquad (4)$$

Also, if it were desired to expand the network determinant as a function of some capacitive element, C_2, the equation for the principal determinant would be:

$$\Delta = \frac{\Delta_{2'} + sC_2\Delta_{2'}}{sC_2} \tag{5}$$

where the primed and unprimed subscripts of the determinants convey the same meaning as before. Each of the equations 4 and 5 has two common properties:

(a) The determinant in which the magnitude of the storage element is set equal to infinity is multiplied by the product of the magnitude of that storage element and the variable s.

(b) The determinant in which the element value is set equal to zero is multiplied by unity.

Equations 4 and 5 may be employed in the expansion of the network determinant of an electric network with several storage elements. For example, suppose that it is desired to expand the network determinant of a network with four storage elements L_1, C_2, L_3, and C_4. The expansion would be:

$$\begin{aligned}\Delta = \frac{1}{s^2C_2C_4}\,[&s^4L_1C_2L_3C_4\Delta_{12'34'}+s^3(L_1L_3C_4\Delta_{1234'}+C_2L_3C_4\Delta_{1'2'34'}\\ &+L_1C_2C_4\Delta_{12'3'4'}+L_1C_2L_3\Delta_{12'34})+s^2(L_1L_3\Delta_{1234}+L_1C_2\Delta_{12'3'4}\\ &+L_1C_4\Delta_{123'4'}+C_2L_3\Delta_{1'2'34}+C_2C_4\Delta_{1'2'3'4'}+L_3C_4\Delta_{1'234'})\\ &+s(L_1\Delta_{123'4}+C_2\Delta_{1'2'3'4}+L_3\Delta_{1'234}+C_4\Delta_{1'23'4'})+\Delta_{1'23'4}].\end{aligned} \tag{6}$$

There are several facts which become apparent in the expansion of the network determinant as a function of the storage elements of the network. It is observed that:

(1) The degree of s in each term of the polynomial representation of Δ is exactly the same as the number of storage elements which are set equal to infinity to obtain the coefficients of the term.

(2) In each successive step that the network determinant is expanded as a function of a particular storage element, one set of the products of the storage elements will contain all possible combinations of products which include that storage element, and a second set of the products of storage elements will contain all possible combinations of products which exclude that particular element. Each of the resulting products of the storage elements is multiplied by the determinant of a network which is derived from the original network by setting all of the storage elements included in the product equal to infinity and the remaining storage elements equal to zero.

(3) In the observations of (1) and (2) above, it is seen that the terms comprising each coefficient of the polynomials of equation 6 represents all possible combinations of products of the same number of storage

elements as the exponent of s by which the coefficient is multiplied. Each of the coefficients is the sum of the products of each combination and the determinant of the network derived in the manner described in observation (2).

(4) If the principal determinant is expanded as a function of all of the storage elements in the network the remaining determinants in the expansion of that determinant correspond to purely resistive networks. For this reason the networks which remain will be called "residual networks."

In the expansion of the principal determinant, the values of the storage elements are arranged as products which will be multiplied by the determinant of a residual network found by setting the inductances and the capacitances of the given network equal to the extreme values of zero or infinity in a systematic manner. Furthermore, it becomes apparent that it is necessary to use all possible combinations of the storage elements set to the extreme values of zero and infinity to derive the required set of residual networks from the original network. Therefore, the original network may be reconstructed as a skeleton network in which the storage elements are replaced by switches. For example, the network of Figure 1a may be visualized as being "extremized"[3] to a network with switches shown in Figure 1b.

Suppose that it is desired to calculate the principal determinant of the network of Figure 1a. In accordance with the observations (1) through (4), all possible combinations of products of the storage elements should be determined first; and then each residual network, which corresponds to the determinant multiplied by a particular product of storage element values, is derived. For example, the residual networks of Figures 2a through 2b are the residual networks necessary to find the coefficients of the denominator functions. It is at this juncture that Feussner's method or Kirchhoff's rules are employed to calculate the determinants of each residual network. Suppose that it is desired to calculate the coefficient of s^2. The residual graphs of Figure 2b through 2d are employed in this calculation. The expression for the coefficient of s^2 in the denominator of the network of Figure 1a will be:

$$b_2 = (L_1C_2[(R_4+R_6)(R_2+R_3+R_5)+(R_3+R_2)R_5)]+C_2L_3[(R_5+R_6)[R_1+R_2+R_4]+R_4(R_1+R_2)]/\Delta_2 \quad (7)$$

A phenomenon which exists in Figure 2c needs explaining. The network determinant is zero. This fact, which will be demonstrated, occurs because part of the network is electrically isolated from the remainder of the network and, as a result, some nodes in that network are electrically isolated points. In the network of Figure 3, there are m branches terminating at node P. The meshes of the network will be chosen in a manner as shown on the diagram. Suppose that each of the terminating branch elements are removed from the network one at a time until there remains only one branch m, terminating a P. This is tantamount to cancelling each row and column in the principal determinant. Since all rows and columns of the determinant are removed the determinant value will be 1. When the branch m is removed from the network, the determinant must have the value zero, since there are no columns or rows to collapse in the remaining determinant. This result may be stated in the following manner: When a part of an electrical network is isolated electrically from the remainder of the network, the value of the principal determinant corresponding to the network is zero.

One other network configuration will always give a determinant value of zero. In Figure 4, there is depicted an electric network with m parallel elements between node-points a' and b'. If any one element is set equal to zero, the node points a' and b' are superposed upon each other resulting in a network whose determinant will, according to Feussner's method, be:

$$\Delta = Z_a..Z_b.......Z_m \quad \Delta_N \tag{8}$$

where the Z's are the impedance of the branches between the node points and Δ_N is the determinant of the network N. If any one of the remaining impedances is set equal to zero, the value of the resultant determinant, according to equation 8, will be zero. The preceding observation may be interpreted in the following manner: "The principal determinant of a network with two or more short-circuited elements between any two node points will have the value of zero". The ideas of electrically isolated points and double short-circuits between two node points of a network obviously are aids in the computation of the coefficients of admittance functions from the residual networks.

CALCULATION OF THE NUMERATOR

It is seen in previous paragraphs that the determinants in equations 2 and 3 were related to each other. Each determinant in equation 3 is a co-factor of a corresponding determinant in equation 2. This fact may be employed in the following way: Consider the expansion of the principal determinant of a network with four storage elements in equation 6. Suppose that it is desired to calculate the current response, I_k, in the kth mesh of the given network to an excitation in the ith mesh, E_i. In this case, the equation for the expansion of the co-factor in the numerator of the current response function, Δ_{ik}, will be:

$$\frac{1}{s^2C_2C_4}[L_1C_2L_3C_4\Delta_{ik12'34'}s^4+(L_1C_2L_3\Delta_{ik12'34}+L_1C_2C_4\Delta_{ik12'3'4'}$$
$$+L_1C_3C_4\Delta_{ik1234'}+C_2L_3C_4\Delta_{ik1'2'34'})s^3+(L_1C_2\Delta_{ik12'3'4}$$
$$+L_1L_3\Delta_{ik1234}+L_1C_4\Delta_{ik1'23'4}+C_2L_3\Delta_{ik1'2'34}+C_2C_4\Delta_{ik1'2'3'4'}$$
$$+L_3C_4\Delta_{ik1'234'})s^2+(L_1\Delta_{ik123'4}+C_2\Delta_{ik1'2'3'4}+L_3\Delta_{ik1'234}$$
$$+C_4\Delta_{ik1'234'})s+\Delta_{ik1'23'4}] \qquad (9)$$

where the **subscript** ik after each symbol Δ indicates that the ith row and kth column in the **principal** determinant are collapsed. If equations 6 and 9 are **compared to each other, several facts** become apparent:

(1) The denominator in each equation is a single term which contains a product of the magnitude of all of the capacitive storage elements in the network. When equation 9 is divided by equation 6, the denominator of equation 6 will cancel the denominator of equation 9. As a result, it may be said that the numerators of the determinants, Δ_{ik}, and Δ, are identical with the numerator and denominator of the admittance function I_k/E_i.

(2) Each determinant or minor in the expansion for the determinant, Δ_{ik}, can be interpreted as being the co-factor of some determinant in the expansion of the principal determinant, Δ. A determinant in Δ_{ik} and its corresponding determinant in Δ are multiplied by identical products of storage elements. For example, Δ_{ik1234} and Δ_{1234} are each multiplied by the products L_1L_3.

(3) The ratio of corresponding determinants in the expansion of Δ_{ik} and Δ is identical to the ratio of the response in the kth mesh of a residual network to the excitation in the ith mesh.

In order to calculate each of the determinants in Δ_{ik}, it is necessary to use the same techniques that were used in calculating the numerator of admittance functions by Feussner's method[1,2] or Kirchhoff's rules[4].

Since each determinant in Δ_{ik} is simply a co-factor of a corresponding determinant in Δ, when Δ_{ik} and Δ are expanded in the manner specified to produce equations 7 and 9, it is sufficient to use the concept of path currents to calculate the function, ρ_f. For example, when it is desired to calculate the determinant, Δ_{ik1234}, it is first necessary to trace the circuit of each of the path currents in the residual network which corresponds to the determinant Δ_{1234} and contains both the branch in which I_k flows and the source E_i. Then each function, ρ_f, corresponding to each path current J_f is calculated. The algebraic sum of the ρ's is the determinant, Δ_{ik1234}.

AN EXAMPLE OF THE CALCULATION OF THE NUMERATOR FUNCTION

In the example of Figure 1a, suppose that it is desired to calculate the current response, I_b, in branch 2. The path currents J_1 and J_2, which flow through the branch containing the resistance R_2 and the branch containing the source e.m.f., E_a, are traced in Figure 1c. First the residual networks are inspected to find the networks which contain complete paths. In this case, the networks of Figures 2d and 2f are the only networks in which the path currents J_1 and J_2 exist. As a result, it is necessary to calculate only the co-factors $\Delta_{ab1'2'3}$ and $\Delta_{ab1'2'3'}$ to find the current response function, I_b. Applying Feussner's method[1,2] to the networks of 2d and 2f, it is seen that:

$$\Delta_{ab1'2'3} = R_4+R_5+R_6 \qquad (10)$$

$$\Delta_{ab1'2'3'} = [R_3[R_5+R_6]+R_5R_6]+R_3R_4. \qquad (11)$$

The equation for the numerator function Δ_{ab} will be:

$$\Delta_{ab} = s^2L_3C_2(R_4+R_5+R_6)+sC_2[R_3(R_5+R_6)+R_5R_6+R_3R_4]. \qquad (12)$$

A helpful aid in calculating the numerator of a current response function is deduced from an inspection of the construction of each residual network in Figure 2. The current I_b can exist in any of the structures only if the inductance L_1 is zero and the capacitance C_2 is infinite in the process of extremization. The extreme value of L_3 has no bearing on whether or not current flows in branch b. This fact means that the paths including the potential source and the specified branch must be complete circuits. With this fact in mind, it may be seen that the networks which are excluded in the computation of the numerator of the current response, I_b, are residual networks in which L_1 is "extremized" to infinity and in which C_2 is "extremized" to zero.

It is also discerned that the value of the coefficients may be zero; and that there is no danger in allowing the maximum degree of s in the numerator and denominator to each be equal to the number of storage elements in the network.

AN EXAMPLE

Application of techniques described in the previous sections will be applied to the unbalanced bridge in Figure 5a. Following the procedures described in the preceding section, the residual networks are derived from the original network in Figure 5a by obtaining all possible combinations of the extremum values of the storage elements. The resulting networks are shown in Figures 4b and 5b and the value of the corresponding determinant of each network is tabulated below it. The co-factors of the determinant required for finding I_a or I_b are indicated by the subscripts a and b, respectively. It is interesting to note that the network of Figure 5d does not enter into the calculations for the currents I_a and I_b (although its network determinant has a value other than zero). While the networks of Figures 5c and 5e appear to provide a path for current I_b, there is a short circuit across the path and as stated previously, networks of this type can be ignored in the calculation of the numerator of the response I_b. On the basis of these observations, it is seen that networks of Figures 5b and 5f are the only ones which enter into the calculation of the numerator of the response I_b. The expressions for the denominator and the numerators of the currents I_a and I_b are presented on the following page.

$$\Delta = [s^3L_1C_2L_3(R_1+R_2+R_3)+s^2(L_3C_2R_2(R_1+R_3)+L_1L_3+L_1C_2R_3(R_1+R_2)) +s(C_2R_1R_2R_3+L_1(R_1+R_2)+L_3(R_1+R_3))+(R_1+R_3)R_2+R_1R_3]/sC_2 \qquad (13a)$$

$$\Delta_{aa} = [s^3L_1C_2L_3+s^2(C_2L_3R_2+L_1C_2R_3)+s(C_2R_2R_3+L_1+L_3)+(R_2+R_3)]/sC_2 \qquad (13b)$$

$$\Delta_{ab} = [-s^3L_1C_2L_3+s(C_2R_2R_3)]/sC_2 \qquad (13c)$$

Thus the equations for I_a and I_b will be:

$$I_a = \frac{\Delta_{aa} E_a}{\Delta} \qquad (14a)$$

$$I_b = \frac{\Delta_{ab} E_a}{\Delta} \qquad (14b)$$

It must be kept in mind that each of the final expressions above conforms to the results of equations 8 and 9.

CONCLUSIONS

The method of residual networks developed in this paper is a systematic approach to the analysis of electric networks.

The method has some advantages over other methods of topological analysis. For example, there are several means of verifying errors in computation. Also, an insight is gained in knowing the effect of circuit elements upon coefficients of network functions by inspection of the network structure. It has been noted that each coefficient of the network functions may be calculated independently of the other coefficients in the network functions by inspecting the proper residual networks. As a matter of fact, there is no need to worry about the degree of the polynomials in the numerator and denominator of network functions because a quick inspection of the residual networks will show whether any coefficient of the polynomials is zero.

There are several interesting sidelights in the analysis of networks by use of the methods of residual networks. The number of residual networks required to find the coefficient of a given power of s in the network determinant is identical to the corresponding coefficient in the polynomial expansion of $(s+1)^n$ where n is the number of storage elements in a network. Therefore, a quick check tells an investigator if he has at least found the correct number of residual networks. The current paths under given conditions are known and, as a result of this, the effect of circuit storage elements upon each coefficient in the response function is readily ascertained.

If the well known methods of topological analysis derived by Kirchhoff[4], Maxwell[5], and Feussner[1,2], and Percival[6,7] are applied to electric networks the equations of the resulting network functions would be in terms of the branch impedances. This fact is appreciated more when it is noted that the methods of the early investigators were developed for application to networks with purely resistive elements and that the resistances in a branch can be lumped into one resistance. Whenever any one of the conventional methods of topological analysis is employed in the analysis of electric networks containing three kinds of elements, the task of performing the algebraic manipulations and collecting terms containing the same power of s together remains to be accomplished. This task may be tedious even for the simple networks and, as stated previously, the methods do not contain the means for verifying errors in computations which are contained in the method of residual networks.

BIBLIOGRAPHY

1. W. Feussner, Uber Stromverzweigung in Netzformigen Leitern, Annalen der Physik, Volume 9, Fourth Series, 1902, pp. 1305-1329.
2. W. Feussner, Zur Berechnung der Stromstarke in Netzformigen Leitern, Annalen der Physik, Volume 15, Fourth Series, 1904, pp. 365-374.
3. Z. Hennyey, Linear Electric Circuits, Addison-Wesley, Reading, Massachusetts, 1962.
4. G. Kirchhoff, Uber die Auflosung der Gleichungen, auf welche man bei der Untersuchungen der Linearen Verteilung Galvanisher Strome gefuhrt wird, Poggendorf's Annalen der Physik, Volume 72, 1847, pp. 497-508.
5. J. C. Maxwell, Electricity and Magnetism, Volume 1, Clarendon Press, Oxford, 1892.
6. W. S. Percival, Solution of Passive Electrical Networks by Means of Mathematical Trees, Journal Institute of Electrical Engineers, London, Volume 100, Part III, 1953, pp. 143-150.
7. W. S. Percival, Improved Matrix and Determinant Methods Solving Networks, Journal Institute of Electrical Engineers, London, Volume 101, Part IV, 1954, pp. 258-264.

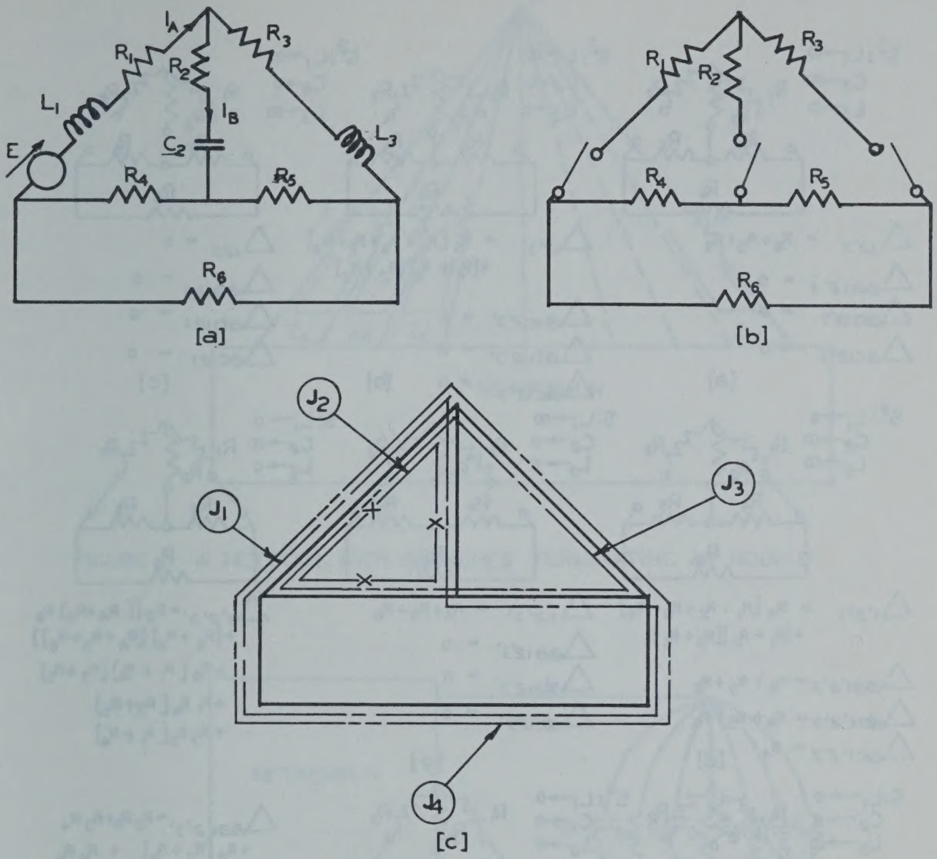
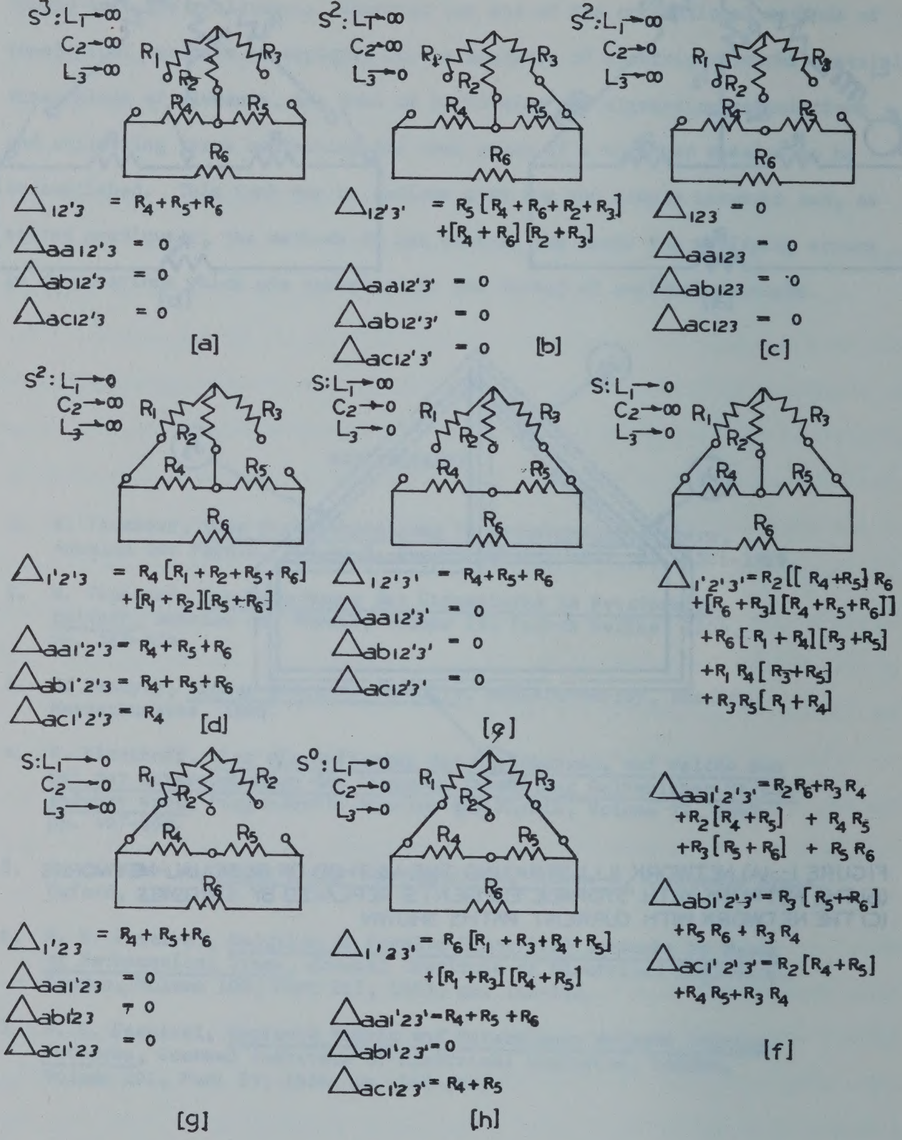

I_A
R_1
R_2
R_3
L_1
I_B
E
C_2
L_3
R_4
R_5
R_6
[a]
R_1
R_2
R_3
R_4
R_5
R_6
[b]
J_2
J_1
J_3
J_4
[c]

FIGURE 1--(A) NETWORK ILLUSTRATING THE METHOD OF RESIDUAL NETWORKS
(B) THE NETWORK WITH STORAGE ELEMENTS REPLACED BY SWITCHES
(C) THE NETWORK WITH CURRENT PATHS SHOWN



$S^3: L_1 \to \infty$, $C_2 \to \infty$, $L_3 \to \infty$

$\Delta_{12'3} = R_4 + R_5 + R_6$

$\Delta_{aa12'3} = 0$

$\Delta_{ab12'3} = 0$

$\Delta_{ac12'3} = 0$

[a]

$S^2: L_1 \to \infty$, $C_2 \to \infty$, $L_3 \to 0$

$\Delta_{12'3'} = R_5[R_4 + R_6 + R_2 + R_3] + [R_4 + R_6][R_2 + R_3]$

$\Delta_{aa12'3'} = 0$

$\Delta_{ab12'3'} = 0$

$\Delta_{ac12'3'} = 0$

[b]

$S^2: L_1 \to \infty$, $C_2 \to 0$, $L_3 \to \infty$

$\Delta_{123} = 0$

$\Delta_{aa123} = 0$

$\Delta_{ab123} = 0$

$\Delta_{ac123} = 0$

[c]

$S^2: L_1 \to 0$, $C_2 \to \infty$, $L_3 \to \infty$

$\Delta_{1'2'3} = R_4[R_1 + R_2 + R_5 + R_6] + [R_1 + R_2][R_5 + R_6]$

$\Delta_{aa1'2'3} = R_4 + R_5 + R_6$

$\Delta_{ab1'2'3} = R_4 + R_5 + R_6$

$\Delta_{ac1'2'3} = R_4$

[d]

$S: L_1 \to \infty$, $C_2 \to 0$, $L_3 \to 0$

$\Delta_{12'3'} = R_4 + R_5 + R_6$

$\Delta_{aa12'3'} = 0$

$\Delta_{ab12'3'} = 0$

$\Delta_{ac12'3'} = 0$

[g]

$S: L_1 \to 0$, $C_2 \to \infty$, $L_3 \to 0$

$\Delta_{1'2'3'} = R_2[[R_4 + R_5]R_6 + [R_6 + R_3][R_4 + R_5 + R_6]] + R_6[R_1 + R_4][R_3 + R_5] + R_1R_4[R_3 + R_5] + R_3R_5[R_1 + R_4]$

$\Delta_{aa1'2'3'} = R_2R_6 + R_3R_4 + R_2[R_4 + R_5] + R_4R_5 + R_3[R_5 + R_6] + R_5R_6$

$\Delta_{ab1'2'3'} = R_3[R_5 + R_6] + R_5R_6 + R_3R_4$

$\Delta_{ac1'2'3'} = R_2[R_4 + R_5] + R_4R_5 + R_3R_4$

[f]

$S: L_1 \to 0$, $C_2 \to 0$, $L_3 \to \infty$

$\Delta_{1'23} = R_4 + R_5 + R_6$

$\Delta_{aa1'23} = 0$

$\Delta_{ab1'23} = 0$

$\Delta_{ac1'23} = 0$

[g]

$S^0: L_1 \to 0$, $C_2 \to 0$, $L_3 \to 0$

$\Delta_{1'23'} = R_6[R_1 + R_3 + R_4 + R_5] + [R_1 + R_3][R_4 + R_5]$

$\Delta_{aa1'23'} = R_4 + R_5 + R_6$

$\Delta_{ab1'23'} = 0$

$\Delta_{ac1'23'} = R_4 + R_5$

[h]

FIGURE 2--(A) THROUGH (H) RESIDUAL NETWORKS DERIVED FROM THE NETWORK OF FIGURE I

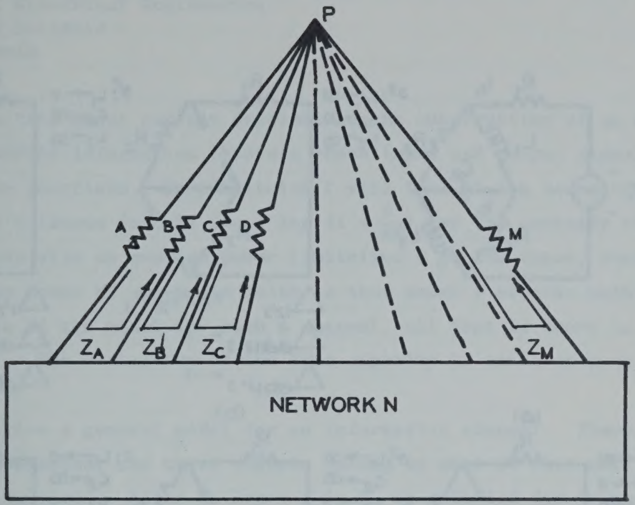

P
A
B
C
D
M
Z_A
Z_B
Z_C
Z_M
NETWORK N

FIGURE 3--A NETWORK WITH BRANCHES TERMINATING AT NODE P

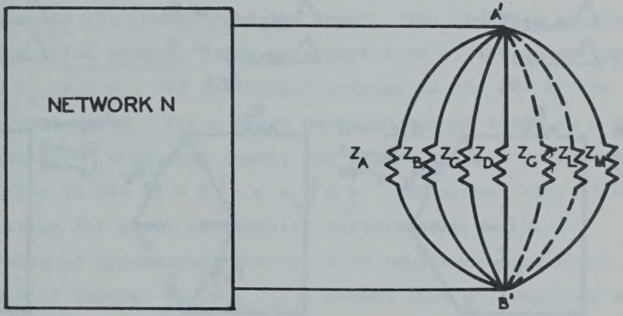
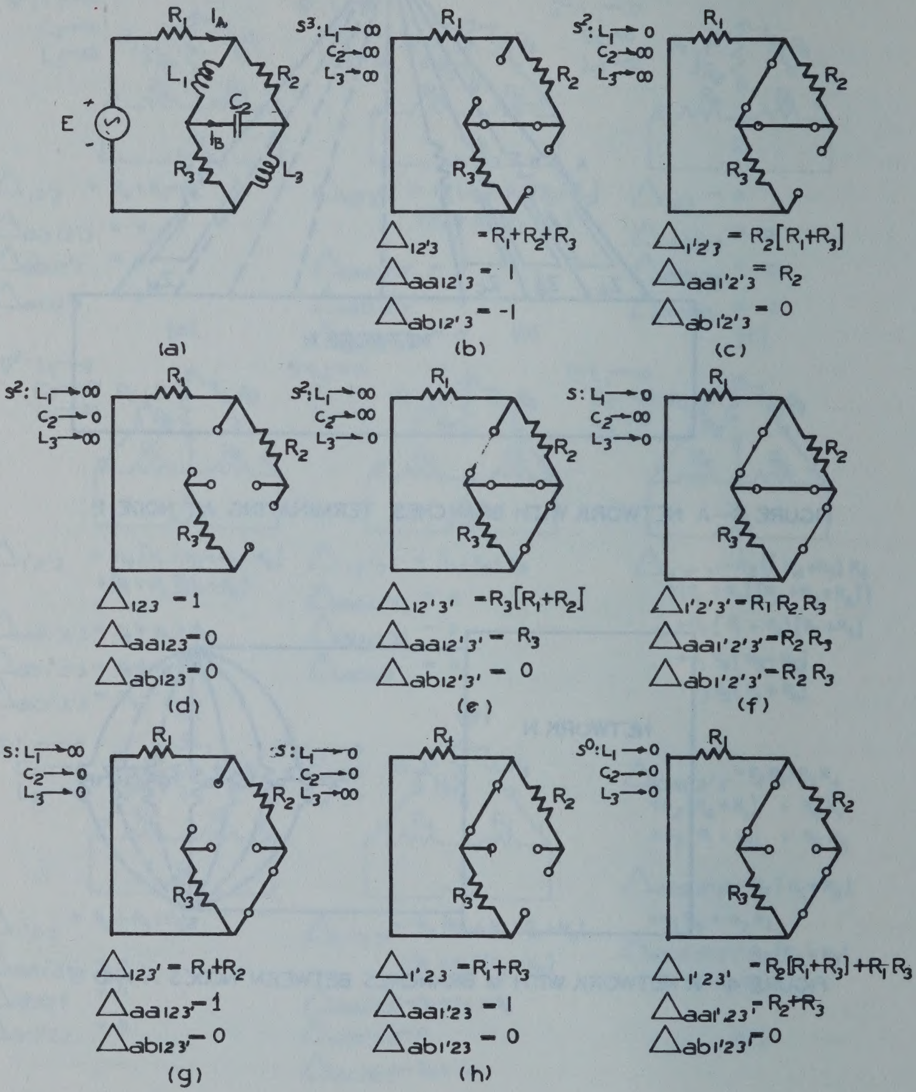

NETWORK N
A'
B'
Z_A
Z_B
Z_C
Z_D
Z_G
Z_L
Z_M

FIGURE 4--A NETWORK WITH M BRANCHES BETWEEN NODES A' AND B'


R1
I_A
L1
R2
C2
E
+
-
I_B
R3
L3
S³: L1 → ∞
C2 → ∞
L3 → ∞
S²: L1 → 0
C2 → ∞
L3 → ∞
S²: L1 → ∞
C2 → 0
L3 → ∞
S²: L1 → ∞
C2 → ∞
L3 → 0
S: L1 → 0
C2 → ∞
L3 → 0
S: L1 → ∞
C2 → 0
L3 → 0
S: L1 → 0
C2 → 0
L3 → ∞
S⁰: L1 → 0
C2 → 0
L3 → 0

(a)

$\Delta_{12'3} = R_1+R_2+R_3$

$\Delta_{aa12'3} = 1$

$\Delta_{ab12'3} = -1$

(b)

$\Delta_{1'2'3} = R_2[R_1+R_3]$

$\Delta_{aa1'2'3} = R_2$

$\Delta_{ab1'2'3} = 0$

(c)

$\Delta_{123} = 1$

$\Delta_{aa123} = 0$

$\Delta_{ab123} = 0$

(d)

$\Delta_{12'3'} = R_3[R_1+R_2]$

$\Delta_{aa12'3'} = R_3$

$\Delta_{ab12'3'} = 0$

(e)

$\Delta_{1'2'3'} = R_1 R_2 R_3$

$\Delta_{aa1'2'3'} = R_2 R_3$

$\Delta_{ab1'2'3'} = R_2 R_3$

(f)

$\Delta_{123'} = R_1+R_2$

$\Delta_{aa123'} = 1$

$\Delta_{ab123'} = 0$

(g)

$\Delta_{1'23} = R_1+R_3$

$\Delta_{aa1'23} = 1$

$\Delta_{ab1'23} = 0$

(h)

$\Delta_{1'23'} = R_2[R_1+R_3]+R_1 R_3$

$\Delta_{aa1'23'} = R_2+R_3$

$\Delta_{ab1'23'} = 0$

FIGURE 5-(A)THROUGH(I) A BRIDGE NETWORK WITH ITS RESIDUAL NETWORKS AND DETERMINANTS

TIME-CONTINUOUS INFORMATION CHANNELS

R. B. ASH
Department of Electrical Engineering
University of Illinois
Urbana, Illinois

I am going to talk about certain problems in the construction of an adequate model to describe information channels whose input and output signals are continuous time functions. In particular I will look at the assumptions underlying Shannon's famous formula $C = W \log (1 + \frac{S}{NW})$ for the capacity of a band-limited channel with an average power limitation S on the input, wideth W, and noise power N. My basic point is that under a precise mathematical formulation of the model for such a channel, all that is known is that $C \geq W \log (1 + \frac{S}{NW})$. The evaluation of the true capacity is still an open problem.

Let me first give a general model for an information channel. There are signals to be communicated, and these signals belong to what we will call the set of allowable inputs G_i. After an input x in G_i is selected for transmission the channel acts on the input to produce an output belong to a set of possible outputs G_o. The mechanism of the channel is described not in a deterministic way but in accordance with a probability distribution $p(y|x)$ defined on G_o for each x in G_i. A channel may be specified by giving the sets G_i and G_o along with the distribution $p(y|x)$. After the output is received, a decision is made as to the identity of the input. The decision process may be described by a decision scheme, which is a partition of the output space G_o into disjoint sets A_x, $x \in G_i$. If the output belongs to the set A_x the receiver declares that the input was x. For a given decision scheme there is a probability of error associated with each input; the probability of error when the input x is transmitter is $p(e\ x) = P\left\{y \in A_x \mid x\right\}$. The probability of error may be calculated using the given probability distribution $p(y\ x)$.

The basic problem of information theory is to construct good codes. A code is a finite set of inputs $\left\{x_1, x_2, \ldots\right\}$ chosen from G_i, together with a decision scheme determined by sets $A_{x_1}, A_{x_2}, \ldots$. The x_i are called code words. The idea is to find codes with a large number of code words while at the same time making the probability of error small. By the probability of error of a code we mean the maximum of the numbers $p(e|x)$ where x ranges over all code words. Thus the problem is to simultaneously work for a large information handling ability and a high reliability. These two requirements are in conflict; however the notion of a channel capacity, to be defined shortly, gives precise results concerning the interplay between the two objectives.

Let me specialize to the time-discrete zero-memory Gaussian channel considered by Shannon in his original paper. Here the allowable inputs are real n-vectors $\underline{x} = (x_1, .., x_n)$ satisfying the "average power limitation" $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq S$. For a given input $\underline{x}$, the output is an n-vector $\underline{y} = \underline{x} + \underline{z}$ where the components of $\underline{z} = (z_1, .., z_n)$ are independent Gaussian random variables with zero mean and variance N. Actually we have defined not a single channel but a collection of channels, one for each positive integer n.

A number R is called a permissible rate of transmission for the above channel if for each n there is a code consisting of $[2^{nR}]$ code words of length n and a probability of error $p_n(e)$ which approaches zero as $n \rightarrow \infty$. In other words, by making n sufficiently large we can achieve an arbitrarily small probability of error while at the same time maintaining the transmission rate R. To justify that the number R may be interpreted as a transmission rate, note that a source of information which produces R binary digits per second will produce nR digits in n seconds; hence in n seconds the source will produce any one of 2^{nR} possible messages. If the information produced by the source is to be transmitted through a channel which transmits 1 component of a code word per second, then in order to process the information produced by the source in the same period of time i.e. n seconds, we need 2^{nR} code words of length n.

The channel capacity C is defined as the maximum (or least upper bound) of all permissible transmission rates. In his original paper, Shannon proved that $C = 1/2 \log_2 (1 + \frac{S}{N})$.

My work is concerned with a generalization of the above channel model to situations where the inputs and outputs are time functions rather than sequences. The model which I consider is the following: (the model was originally introduced by Fortet[4] and Bethoux[3]). Given a Gaussian noise process n(t) with spectral density $N(\omega)$ (not necessarily band-limited). For a given positive number T, consider the class of functions s(t) whose Fourier transforms $S(\omega)$ are such that:

1. $S(\omega) = 0$ whenever $N(\omega) = 0$, and
2. $\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{|S(\omega)|^2}{N(\omega)} \, dw \leq KT$ where K is a fixed positive constant.

The integrand is defined to be zero if $N(\omega) = 0$. The constraint (2) is in a sense a natural generalization of the average power limitation for the time-discrete zero memory channel, as we shall see later.

The allowable inputs are truncated versions $s_T(t)$ of signals in the above class, i.e. $s_T(t) . s(t)$, $0 \leq t \leq T$; $s_T(t) = 0$ elsewhere, where s(t) satisfies (1) and (2) above. The output is the sum of the input and the truncated noise $n_T(t)$: $y(t) = s(t) + n(t)$, $0 \leq t \leq T$; $y(t) = 0$ elsewhere. Analogous to the time-discrete case, a number R is called a permissible rate of transmission if for each T there is a code for the above channel consisting of $[2^{RT}]$ inputs of duration T, with a probability of error $p_T(e)$

which approaches zero as $T \rightarrow \infty$. The channel capacity is the least upper bound of all permissible transmission rates.

The capacity of the channel described above turns out to be $\frac{K}{2} \log_2 e$ (See references 1 and 2). If we specialize the result to the case where the noise is flat band-limited, we do not obtain Shannon's formula.

Let us consider the band-limited case. We have

$$N(\omega) = \frac{N}{2}, \ -2\pi W \leq \omega \leq 2\pi W$$

$$= 0 \text{ elsewhere}$$

If we take $K = \frac{2S}{N}$, the constraints on the input are first that the input signals be limited to the same frequency band as the noise, and second that

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} |S(\omega)|^2 \, d\omega \leq ST$$

or, using Plancherel's theorem,

$$\frac{1}{T} \int_{-\infty}^{\infty} s^2(t) \, dt \leq S$$

i.e. the classical "average power" limitation. The capacity is $C = \frac{K}{2} \log_2 e = \frac{S}{N} \log_2 e$ which is the limit as $W \rightarrow \infty$ of Shannon's expression $W \log_2 (1 + \frac{S}{NW})$.

The basic reason for the discrepance between the above results and those of Shannon is that the process of taking a band limited signal, truncating it, adding it to a section of band-limited noise, observing the result and then making a decision as to the identity of the input is not adequate to describe a band-limited channel. Effectively we are time-limiting the code words, and thus it is not surprising that the channel capcity agrees with Shannon's formula for the infinite bandwidth case. Any realistic model for a band-limited channel must take into account the fact that if a section of a band-limited signal s(t) is transmitted during the interval [0,T], the portion of s(t) for t outside [0,T] remains to interfere with signals to be transmitted during the intervals [T,2T], [2T,3T] etc. Strictly speaking there will be interference with signals transmitted during the intervals [-T,0], [-2T,-T], as well. Thus in a given interval, say [0,T], the actual output is $s^*(t) + n(t)$, where

$$s^*(t) = s(t) + \sum_{\substack{k=-\infty \\ R\neq 0)}}^{\infty} s_k(t-kT)$$

where $s_k(t-kT)$, $kT \leq t \leq (k+1)T$, is transmitted during the interval $[kT,(k+1)T]$.

Shannon, in his original paper, solved the interference problem by using the sampling theorem. He used a decoding procedure based on observation of the output at discrete sampling instants, and he restricted the class of allowable inputs to those having only a finite number of non-zero samples. If the non-zero samples of one function are restricted to the interval [0,T] and the non-zero samples of another function to the interval $[T, 2T]$, there will be no interference between the two functions.

Shannon's results can be interpreted as a rigorous proof that the capacity C^* of the band-limited shannel (with interference taken into account) is at least $W \log_2 \left(1 + \frac{S}{NW}\right)$. Since Shannon assumed a specific decoding procedure and a specific way of choosing code words, it is not possible to conclude from his results that $C^* = W \log_2 \left(1 + \frac{S}{NW}\right)$; it is conceivable that Shannon's procedure is not optimal. Thus the evaluation of C^* is still an open problem. All we know for sure is that

$$C^* \leq C = \frac{S}{N} \log_2 e.$$

We can also oberve that if the bandwidth is infinte, we can solve the interference problem by using time-limited code words. Then Shannon's problem is the same as ours and $C^* = C = \frac{S}{N} \log_2 e$.

References

1. Ash, R. B.: "Capacity and Error Bounds for a Time-Continuous Gaussian Channel," Information and Control, March 1963, pp. 14-27.

2. Ash, R. B.: "Further Discussion of a Time-Continuous Gaussian Channel," Information and Control, to be published.

3. Bethoux, P.: "Test et estimations concernant certaines functions aleatoires en particulier Laplaciennes," Annales de l'Institut Henri Porncare, 1962, Vol. 27, pp. 255-322.

4. Fortet, R.: "Hypothesis Testing and Estimation for Laplacian Functions," Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1961, Vol. I, pp. 289-305.

OPTIMAL CONTROL OF MULTIPLICATIVELY PERTURBED MARKOV STOCHASTIC SYSTEMS*

WILL GERSCH and FRANK KOZIN
Aeronautical and Engineering Sciences
Purdue University
Lafayette, Indiana

ABSTRACT

The problem of optimal control of independent increment multiplicatively perturbed stochastic systems subject to a quadratic loss function is treated.

The partial differential dynamic transition equations for Markov processes are developed by a dynamic programming approach and solved by a series solution technique. The development is applicable to both multiplicative and additive independent increment perturbations of a deterministic system and to completely stochastic systems. The optimal control is linear and the optimally controlled linear system is stable in the mean square sense. The work is an extension of Florentin's efforts in optimal control of independent increment additively perturbed stochastic systems.

I. Introduction

Linear stochastic sequences may be dichotomized into two distinct classes. The first is the class in which a deterministic system is subject to additive external stochastic influences and has been referred to as the additive class. The second class is that in which the internal mechanism of the system varies stochastically. This class is referred to as the multiplicative class. These classes have been discussed by Bellman.[1] [**]

In this paper, we treat the problem of optimally controlled independent increment[+] multiplicatively perturbed stochastic sequences subject to a quadratic loss function.

The dynamic programming approach of Bellman[2] is used to develop the dynamic transition equations for Markov processes. The development is applicable to both multiplicative and additive independent increment perturbations of a deterministic system and to more general stochastic systems defined by their transition matrices. Earlier, Florentin[3] and Kushner[4] treated the problem of optimum stochastic control, however, their attention was restricted to the additive class of independent increment disturbances.

Consider the controlled n-dimensional Markov diffusion process

---

[*] Presented at the First Allerton Conference on Circuit and Circuit Theory, of The University of Illinois, November 1963. Sponsored in part by National Science Founcation.

[**] Numbers refer to the Bibliography.

[+] The continuous parameter process $\left\{Y_t; t\in T\right\}$ is said to have independent increments if for any n and any choice $t_1 st_2$ .... stn∈T the random variables $Y_{\epsilon 2}- Y_{\epsilon 12}\ldots Y_{tn}- Y_{tn-1}$ are independent.

$\left\{X_t, - \leq t < \infty\right\}$, defined by the stochastic differential equations

$$\dot{X}_t = A(t,X_t) + \dot{\xi}_t X_t + \Theta_t + \dot{\eta}_t, \tag{1.1}$$

where X_t is the value of the state n-vector at time t, A(t,X) is a given n-vector function, ξ_t is an nxn matrix of independent increment processes $(\xi_{ij,t})$, η_t is an independent increment n-vector process (assumed to be independent of ξ_t), and Θ_t is the n-vector control function that is to be determined so as to extremize a performance index.

It is important to note that Eq. (1.1) exists only in a formal way since the indicated derivatives of ξ_t and η_t do not exist in the mean square sense or in fact in any sense. The equation can however be made meaningful if the indicated derivatives are interpreted as differentials. This is carried out explicitly below.

The performance index is taken to be the conditional expectation of a quadratic path integral functional. For the final value problem, the expectation of this functional is given as

$$E\left\{\int_{t_o}^{s} L(X_u, \Theta_u; u)\,du \,\middle|\, X_o, t_o\right\} \tag{1.2}$$

The problem is to minimize this functional over the class of control functions Θ. This minimum, denoted $K(X_o,t_o)$ is given by

$$K(X_o,t_o) = \min_{\Theta} E\left\{\int_{t_o}^{s} L(X_u,\Theta_u;u)\,du \,\middle|\, X_o, t_o\right\} \tag{1.3}$$

The dynamic transition equations for $k(X_o,t_o)$ are derived in the next section from the Bellman Dynamic Programming algorithm.

II. Dynamic Transition Equation

Using the additivity of the expectation operator and the principle of optimality, the minimum of the expected quadratic performance index may be written as

$$K(X_o,t_o) = \min_{\Theta(t_o,s)} E\left\{\int_{t_o}^{s} L\,du \,\middle|\, X_o, t_o\right\} \tag{2.1}$$

$$= \min_{\Theta(t_o,s)} E\left\{\int_{t_o}^{t_o+\Delta t} L\,du + \int_{t_o+\Delta t}^{s} L\,du \,\middle|\, X_o, t_o\right\}$$

$$= \min_{\Theta(t_o,t_o+\Delta t)}\left[E\left\{\int_{t_o}^{t_o+\Delta t} L\,du \,\middle|\, X_o, t_o\right\} + \min_{\Theta(t_o+\Delta t,s)} E\left\{\int_{t_o+\Delta t}^{s} L\,du \,\middle|\, X_o, t_o\right\}\right].$$

The controlled process, X_t, is a Markov process and the optimum policy to be determined is independent of the explicit value of ΔX. Consequently, the second expectation on the right hand side of the last equality in (2.1) may be written as

$$\min_{\Theta(t_o+\Delta t,s)} E\left\{\int_{t_o+\Delta t}^{s} L\ du \,\middle|\, X_o,t_o\right\} = \min_{\Theta(t_o+\Delta t,s)} E\left\{E\left\{\int_{t_o+\Delta t}^{s} L\ du \quad \begin{matrix} X_o+\Delta x, t_o+\Delta t \\ X_o, t_o \end{matrix}\right\} \,\middle|\, X_o,t_o\right\} \tag{2.2}$$

$$= \min_{\Theta(t_o+\Delta t,s)} E\left\{E\left\{\int_{t_o+\Delta t}^{s} L\ du \,|\, X_o+\Delta X, t_o+\Delta t\right\} \,|\, X_o,t_o\right\}$$

$$= E\left\{K(X_o+\Delta X, t_o+\Delta t) \,|\, X_o,t_o\right\}$$

Substituting (2.2) into (2.1), to the first order in Δt we obtain

$$K(X_o,t_o) = \min_{\Theta_o} \left[L(X_o,\Theta_o,t_o)\Delta t + E\left\{K(X_o+\Delta X, t_o,+\Delta t \,|\, X_o,t_o\right\}\right] \tag{2.3}$$

Equation (2.3) is, in fact, the crux of optimal control of Markov sequences. We interpret Eq. (2.3) for the two important independent increment processes, the Wiener process and the Poisson process. Consider first the Wiener process.

A. Wiener Process

The one-dimensional process $\left\{Y_t,\ 0 \le t < \infty\right\}$ is said to be a Wiener process if it is a normally distributed stationary independent increment process satisfying $E\left\{Y_t\right\}$ for $t>0$ and $Y_o = 0$ with probability one. It follows that $E\left\{(Y_t-Y_s)^2\right\} = \sigma^2|t-s|$. Let us assume that the elements of the matrix ξ_t and the components of the vector η_t appearing in (1.1) are Wiener processes. As an equation in differentials, to the first order in Δt, Eq. (1.1) is written

$$\Delta X_t = A(t,X_t)\Delta t + \Delta\xi\, X_t + \Theta_t\Delta t + \Delta\eta_t + 0\ (\Delta t) \tag{2.4}$$

where

$$\Delta X_t = X_{t+\Delta t} - X_t,\ \Delta\xi_t = \xi_{t+\Delta t} - \xi_t,\ \Delta\eta_t = \eta_{t+\Delta t} - \eta_t. \tag{2.5}$$

This gives rise to an n-dimensional Markov diffusion process whose first two derivate moments are defined by

$$\lim_{\Delta t \to o} \frac{1}{\Delta t} E\left\{\Delta X_t \,|X\right\} = b(X_1 t) \tag{2.6}$$

$$\lim_{\Delta t \to 0} \frac{i}{\Delta t} \; E\left\{ (\Delta X_t)(\Delta X_t^T) \,|X \right\} = D(X_1 t)$$

where $b(X_1 t)$ is an n-vector, $D(X_1 t)$ is an nXn matrix and "T" denotes transpose.

From the definition of the Wiener process, it follows that no higher order derivate moments exist and, furthermore, we have explicitly

$$b(X_1 t) = A(X_1 t) + \Theta(t)$$

$$D(X_1 t) = \lim_{\Delta t \to 0} \frac{}{\Delta t} \left[E\left\{ (\Delta \xi_t X_t)(\Delta \xi_t X_t)^T |X_1 t \right\} \right] = \mathcal{S} + \mathcal{J} \tag{2.7}$$

where $\mathcal{S} = [x_i s_{ij} x_j]$, $\mathcal{J}$- $[t_{ij}]$ are nxn matrices. Each element of $\mathcal{S}$ is a quadratic form in x whose coefficients are the variances and co-variances of the elements of the ξ_t matrix. $\mathcal{J}$ is the variance-co-variance matrix for the η_t process. Using the results above, the right hand side of (2.3) is expanded in a Taylor expansion about (X_o, t_o), up to the first order in Δt[(from 2.6)], this expansion corresponds to the second order in ΔX)

$$\begin{aligned} K(X_o,t_o) = \min_{\Theta_o} \Big[& L(X_o,\Theta_o,t_o)\Delta t + K(X_o,t) + \frac{\partial K(X_o,t_o)}{\partial t_o} \Delta t_o \\ & + E\left\{\Delta X_{to}^T |X_o,t_o\right\} \frac{\partial K(X_o,t_o)}{\partial X_o} \\ & + \frac{1}{2}\left[\left(\frac{\partial}{\partial X_o}\right)^T E\left\{ (\Delta X_{to})(\Delta X_{to})^T |X_o,t_o \right\} \frac{\partial}{\partial X_o} \right] K(X_o,t_o) \\ & + O(\Delta t) \Big]. \end{aligned} \tag{2.8}$$

In Eq. (2.8), the differential operators in the expectation do not act on the vector X.

Now we subtract $K(X_o,t_o)$ from both sides of (2.8), divide by Δt, and let Δt approach zero. Equations (2.6) and (2.7) are inserted into (2.8) to yield

$$\frac{-\partial K}{\partial t_o} = \min_{\Theta_o} \left[L(X_o,\Theta_o,t_o) + b^T(X_o,t_o)\frac{\partial K}{\partial X_o} + \frac{1}{2}\left[\left(\frac{\partial}{\partial X_o}\right)^T D(X_o,t_o)\left(\frac{\partial}{\partial X_o}\right)\right] K \right] \tag{2.9}$$

Here too, $\frac{\partial}{\partial X_o}$ does not operate on $D(X_o,t_o)$.

It is convenient now to drop the subscripts "o" in Eq. (2.9) and write (2.9) in terms of the variable $t = s - t_o$, the "time to go". We then have

$$\frac{\partial K}{\partial t} = \min_{\Theta} \left[L(X_1,\Theta,t) + b^T(X_1 T)\frac{\partial K}{\partial X} + \frac{1}{2}\left[\left(\frac{\partial}{\partial X}\right)^T D(X_1 t)\left(\frac{\partial}{\partial X}\right)\right] K \right] \tag{2.10}$$

where the initial condition is $K(X,0) = 0$.

The loss function is assumed to be of the form

$$L(X,\Theta) = X^T QX + \Theta^T B \Theta \tag{2.11}$$

where Q is the "cost of state" matrix, and B is the "cost of control" matrix, each of which may be a function of time.

Application of (2.7) and (2.11) to (2.10) finally gives

$$\frac{\partial K}{\partial T} = \min_{\Theta} \left[X^T QX + \Theta^T B \Theta + (A(X_1 t) + \Theta)^T \frac{\partial K}{\partial X} + \frac{1}{2} \left(\left(\frac{\partial}{\partial X}\right)^T (\mathcal{S} + \mathcal{J}) \left(\frac{\partial}{\partial X}\right) \right) K \right] \tag{2.12}$$

Extremization of Eq. (2.12) is accomplished as usual, by setting the derivatives of (2.12) with respect to the components of Θ, equal to zero. The resultant equation for the optimal control is found to be

$$\Theta^x = -\frac{1}{2} B^{-1} \quad \frac{\partial K}{\partial X} 1 \tag{2.13}$$

B^{-1} is a suitably defined matrix, in the event the inverse does not exist,.

We insert this equation for optimal control into (2.12). The resulting transition equation is

$$\frac{\partial K}{\partial t} = X^T QX - \frac{1}{4}\left(\frac{\partial K}{\partial X}\right)^T B^{-1} \left(\frac{\partial K}{\partial X} + A(X_1 t)^T \quad \frac{\partial K}{\partial X} + \frac{1}{2} \left(\left(\frac{\partial}{\partial X}\right)^T (\mathcal{S} + \mathcal{J}) \left(\frac{\partial}{\partial X}\right) \right) K \tag{2.14}$$

Equation (2.14) is a nonlinear, second order partial differential equation. The second order terms are easily seen to be due to the presence of variances in the process of order 0(Δt) rather than 0(Δt), which is the case for deterministic processes.

B. The Poisson Process

The Process can be defined directly as an independent increment counting process.[5] It is sufficient for our purposes to consider the interval (t,t+Δt), and say that the probability that an impulse having magnitude in the range (Z,Z + Δ Z) occurs in this time interval is given by dF(Z|t).

For the simple Poisson event with magnitude Z = z_o, we can say

$$\text{Prob.} \left\{ Z = z_o \text{ in } (t,t+\Delta t) \right\} = dH(Z-z_o)\big|_{Z=z_o} (\lambda \Delta t + 0(\Delta t)) \tag{2.15}$$

$$\text{Prob.} \left\{ Z = 0 \text{ in } (t,t+\Delta t) \right\} = dH(Z)\big|_{Z=0} (1-\lambda\Delta t + 0(\Delta t))$$

where H is the unit step function. λ is called the mean rate, or intensity, of the process.

Consider again the stochastic difference equation determined by (1.1). Neglecting, for the moment, the additive noise component η_t, we say that with probability dF(Z|t)

$$\Delta X_t = A(X_t, t)\Delta t + ZX_t + \Theta_t \Delta t + 0(\Delta t) \tag{2.16}$$

where Z in (2.15) represents an nxn diagonal matrix of Poisson processes.

Therefore we can return to Eq. (2.3) and with the right hand term to the first order in Δt as

$$E\left\{K(X_o+\Delta X, t_o+\Delta t)\,|X_o, t\right\} = \tag{2.17}$$

$$\int_Z K(X_o+A(X_o,t_o)\Delta t + ZX_o + \Theta\Delta t + 0(\Delta t),\ t+\Delta t)\ dF(Z|t_o)$$

$$=\int_Z\left\{K(X_o+ZX_o,t_o) + \left[(A(X_o,t) + \Theta_o)^T\ \frac{\partial K}{\partial X_o+ZX_o}\ (X_o+ZX_o,t_o) + \frac{\partial K(X_o+ZX_o,t_o)}{\partial t_o}\right]\Delta t\right\}$$

$$\cdot\ dF(Z|t_o).$$

$$= \int_Z\left\{K(X_o+ZX_o,t_o) + \left[(A(X_o,t) + \Theta_o)^T\ \frac{\partial K}{\partial X_o+ZX_o}\ (X_o+ ZX_o,t_o) + \frac{\partial K(X_o+ZX_o,t_o)}{\partial t_o}\right]\Delta t\right\}$$

$$\cdot\ \lambda\ \Delta t dF(Z)$$

$$+\left\{K(X_o,t_o)\left[(A(X_o,t_o) + \Theta_o)^T\ \frac{\partial K(X_o,t_o)}{\partial X_o}\ \frac{\partial K(X_o,t_o)}{\partial t_o}\right]\Delta t\right\}(1-\lambda\Delta t).$$

Using the expression (2.17) in the fundamental eq. (2.3), dividing by Δt, allowing Δt to approach zero, and transforming to the "time to go" variable t one obtains the equation

$$\frac{\partial K}{\partial t} = \min_{\Theta}\left\{L(X,\Theta) + (A(X,t)+ \Theta^T\ \frac{\partial K}{\partial X} + \lambda\int K(X+ZX,\ t)\ dF(Z)-\lambda K\right\} \tag{2.18}$$

Again with $L(X,\Theta) = X^TQX + \Theta^TB\,\Theta$, extremization of (2.18) requires that (2.13) holds for the Poisson case. The integro-differential equation obtained by substituting the equation for optimal control into (2.18) is,

$$\frac{\partial K}{\partial 2} = X^T(QX - \frac{1}{4}\ (\frac{\partial K}{\partial X})^T B^{-1}\ (\frac{\partial K}{\partial X}) + A(X,t)\ \frac{\partial K}{\partial X} + \lambda\int_Z K(X+ZX,t)\ dF(Z) \tag{2.19}$$
$$-\lambda k.$$

This equation is the general transition equation for the optimal control of Markov processes with multiplicative Poisson increments. If η_t, an additive Poisson noise were present in the system (and independent of ξ_t), and if the increments Y of η_t were distributed according to $G(y,t)$ with mean rate λ_y, the resulting expression would be

$$\frac{\partial K}{\partial 2} = X^TQX - \frac{1}{4}\ (\frac{\partial K}{\partial X})^T\ B^{-1}(\frac{\partial K}{\partial X}) + A(X_1t)^T\ \frac{\partial K}{\partial X}$$
$$+ \lambda_Z\int K(X+ZX,t)dF(Z) + \lambda_Y\int K(X+Y,t)dG(Y) - (\lambda_Z+\lambda_Y)K \tag{2.20}$$

III. Some One-Dimensional Illustrations

A. Wiener Process.

We consider the simple specific case given by the linear equation

$$\Delta X_i = aX_t\Delta t + \Delta \xi X_t + \Theta_t\Delta t + 0(\Delta t) \tag{3.1}$$

In Eq. (3.1), a is constant and known, $\left\{ \xi_t, t \in [0,\infty) \right\}$ is the Wiener process with $\left\{ E \quad (\Delta \xi_t)^2 \right\} = \sigma^2\Delta t$

We assume the loss functional to be

$$\int_{t_o}^{S} (X_u^2 + \Theta_u^2)\, du \tag{3.2}$$

Application of (2.12) to the system (3.1), with the loss function (3.2) gives

$$\frac{\partial K}{\partial t} = \min_{\Theta} \left[X^2 + \Theta^2 + (aX + \Theta \;\; \frac{\partial K}{\partial X} + \frac{\sigma^2}{2} \;\; X^2 \;\; \frac{\partial^2 K}{\partial X^2} \right] \;\; . \tag{3.3}$$

The condition for the extremal is

$$\Theta = - \frac{1}{2} \;\; \frac{\partial K}{\partial X} \;\; . \tag{3.4}$$

With condition (3.4), Eq. (3.3) becomes

$$\frac{\partial K}{\partial t} = X^2 + aX \;\; \frac{\partial K}{\partial X} - \frac{1}{4} \;\; \left(\frac{\partial K}{\partial X}\right)^2 + \frac{\sigma^2}{2} \;\; X^2 \;\; \frac{\partial^2 K}{\partial X^2}. \tag{3.5}$$

As first suggested for this problem by Merriam,[6] we assume a series solution for the transition equation in the form

$$K(X_1 t) = \sum_{i=o}^{\infty} k_i(t)X^i \tag{3.6}$$

The initial condition is $K(X_1 0) = 0$ consequently, $k_i(0) = 0$ for all i.

Substitution of (3.6) into (3.5) and subsequent matching in powers, of X, leads us to an infinite set of first order differential equations in the k_i's. By induction, using the given initial conditions, we can conclude that only the k_2 term is non-identically vanishing.

Therefore,

$$\begin{aligned} K(X_1 t) &= k_2(t)X^2 \\ \Theta &= - k_2(t)X^2 \end{aligned} \tag{3.7}$$

The equation for $k_2(t)$ is found from (3.5) to be the Ricatti equation

$$\dot{k}_2 = (\sigma^2 + 2a)k_2 - k_2^2 - k_2^2 + 1, \;\; k_2(0) = 0. \tag{3.8}$$

The solution of this equation is

$$k_2(t) = \lambda_t \left[\frac{1 - \exp(\lambda_- - \lambda_+)\, t}{1 - \frac{\lambda_+}{\lambda_-} \;\; \exp(\lambda_- - \lambda_+)t} \right] \tag{3.9}$$

where

$$\lambda_{\pm} = \frac{\sigma^2}{2} + a \pm \sqrt{\left(\frac{\sigma^2}{2} + a\right)^2 + 1}.$$

Since the optimal control is $\Theta = -\frac{1}{2} k_2(t)X$, for the finite duration control problem, there is engineering interest in the asymptotic behavior of k_2 for t large. For $T-t_0 \gg \lambda_+\lambda_-$, the implementation of the fixed control system $\Theta = -k_2(\infty)X = -\lambda_+ X$ is, in general, a superior engineering decision rather than the alternative of implementing a time variable controller.

Therefore, for the infinite duration control problem, the control solution $\Theta = -\lambda_+ X$ is implemented in the system of Eq. (3.1). The controlled system written in differential equation rather than difference form is

$$\dot{X}_t + (\beta - \xi_t)X_t = 0 \qquad (3.10)$$

where $\beta = \lambda_+ - A > 0$.

An evaluation of the mean square stability of the controlled system is of interest. The system (3.10) is asymptotically stable in the mean square sense if

$$\lim_{t \to \infty} E\left\{X^2(t)\right\} = 0 \qquad (3.11)$$

Equation (3.1) and hence the system mean square stability can be evaluated by the methods proposed by Bogdanoff and Kozin.[7] First it is necessary to obtain the Fokker-Planck equation for the transition density of the Markov process defined by (3.10). The equationmay be obtained, for example, by Bartlett's formalism[8] as

$$\frac{\partial p(X_1 t)}{\partial^2} = \beta \frac{\partial}{\partial X}(X\, p(X_{\cdot} t) + \frac{\sigma 2}{2} \frac{\partial^2}{\partial X^2} (X^2 p(X_1 t)). \qquad (3.12)$$

The expectation is (3.11) may be evaluated by multiplying (3.12) by X^2 and then integrating over the entire range of values of X.[7]

We observe that $\mu_2(t) \equiv E\left\{X^2(t)\right\}$ satisfies the equation

$$\dot{\mu}_2(t) = (-2\beta + \sigma^2)\mu_2(t) \qquad (3.13)$$

$$= -2\sqrt{\left(\frac{\sigma^2}{2} + a\right)^2 + 1}\; \mu_2(t)$$

Obviously, the controlled system is stable in the mean square sense.

B. Poisson Process.

We consider the regulator problem in stochastic difference equation form,

$$\Delta x_t = \Theta_t \Delta t + Z_t X_t. \qquad (3.14)$$

and assume Z_t to be a Poisson process. Under the loss function of Eq. (3.2) the transition equation is

$$\frac{\partial K}{\partial 2} = X^2 - \frac{1}{4}\left(\frac{\partial K}{\partial X}\right)^2 + \lambda \quad dF(Z|X_1 t)K(X_t + Z_t X_t, \ t) - \lambda K(X,t) \tag{3.15}$$

and the optimal control is

$$\Theta = -\frac{1}{2}\left(\frac{\partial K}{\partial X}\right) . \tag{3.16}$$

As in the previous examples we assume the series solution of Eq. (3.6). Again the only non-zero term is the k_2 term. The differential equation for this coefficient is

$$\dot{k}_2 = 1 - \dot{k}_2^2 + \lambda k_2[ZE\ Z + E\ \ Z^2\ \ ]. \tag{3.17}$$

The solution is as in Section III-A. As expected, the form of the solution of the optimal control regulator problem with multiplicative Poisson increments is not different than the solution for the case of multiplicative increments from a Wiener process.

There is, however, an instructive comparison to be made of this example and the additive noise regulator problem considered by Florentin.[9] Florentin considers the regulator governed by the equation

$$\dot{X}_t = \Theta_t + Z_t . \tag{3.18}$$

The form of the dynamic transition equation for this problem is

$$\frac{\partial K}{\partial t} = X^2 - \frac{1}{4}\left(\frac{\partial K}{\partial X}\right)^2 + \lambda \quad dF(Z|X,t)\ K(X + Z_t, t) - \lambda\ K(X,t) \tag{3.19}$$

This differs from Eq. (3.15) only in the form of the perturbation term under the integral sign.

Florentin considers the situation of a nonlinear intensity $\lambda = X^2$ and obtains a solution to (3.19) essentially identical to that obtained for (3.15) with a homogeneous, fixed intensity.

The object of this comparison is to suggest the utility of the multiplicative perturbation stochastic sequence model as a constructive tool in the formulation of models for physical processes. By creating and experimenting with such models we have the means of extending our understanding of natural physical processes. This approach is clearly in its infancy.

IV. The Multivariate Transition Equation.

The transition equations for both of the independent increment models, (2.14) and (2.20) can be satisfied by series expansions of the form

$$K(X,t) = K_o(t) + K_1^T(t)\ X + X^T(K_2(t)X. \tag{4.1}$$

where K_o is a scalar, K_1 an n-vector, and K_2 a symmetric nxn matrix, $K_2 = [k_{ij}(t)]$. The result of substituting (4.1) into the Wiener process example for the general linear model of (1.1), where the matrix is diagonal, yields the set of ordinary differential equations

$$\dot{K}_o = -\frac{1}{4} K_1^T B^{-1} K_1 + \sum_{i,j} k_{tj} t_i, \tag{4.2}$$

$$-\dot{K}_1^T = - K_1^T B^{-1} K_2 + K_1^T A$$

$$\dot{K}_2 = Q - K_2 B^{-1} K_2 + A^T K_2 + K_2 + A + [k_{ij}(t) S_{tj}]$$

with $K_o(o) = 0$, $K_1(o) = 0$, $K_2(o) = 0$

The optimal control is of the form

$$\theta^x = - K_2 X - \frac{1}{2} K_1 \tag{4.3}$$

Similarly, for the Poisson increment problem governed by the transitio equation, (2.20), the set of ordinary differential equations to be solved for the coefficients in the expansion of $K(X_1 t)$ is

$$\dot{K}_o = - \frac{1}{4} K_,^T B^{-1} K_1 \tag{4.4}$$

$$\dot{K}_1 T = - K_1^T B^{-1} K_2 + K_1^T A + \lambda K_1^T E \quad Z_t \quad .$$

$$\dot{K}_2 = Q - K_2 B^{-1} K_2 + A^T K_2 + K_2 A + \lambda E \quad Z_t^T K_2 Z_t + 2\lambda E \quad Z_t^T K_2$$

The optimal control is in the form of Eq. (4.3).

Equations (4.2) and (4.4) demonstrate that only elementary computing is required to solve the optimal control Markov stochastic system problem, for systems with well-behaved time varying parameters. Florentin[3] demonstrated this result for the additive class of independent increment perturbations. The results hereare, therefore, extended to include the independent increment multiplicatively perturbed class. Consequently, we are in a position to solve problems that have not been considered before.

V. Conclusions

We have treated the problem of optimal control of independent increment multiplicatively perturbed stochastic processes. The dynamic programming approach ofBellman is used to develop the partial differential transition equations for Markov processes. These are solved by a series solution technique. The optimal control is linear and, in an example, the optimally controlled linear system was shown to be stable in the mean square sense. This work completes the picture of optimal control of Markov processes started by Florentin on the additive class of problems.

A computational algorithm is now available for a class of problems that have not been solved before. Using an example we have suggested the utility of the multiplicatively perturbed stochastic process as a constructive tool in extending our understanding of physical processes.

The more general problem of optimal control of general linear process multiplicative disturbances remains to be solved.

VI References

1. R. Bellman, "Adaptive Control Processes, A Guided Tour," Princeton Univ. Press, Princeton, New Jersey, 1961, Sec. 9112.

2. R. Bellman, "Dynamic Programming," Princeton Univ. Press, Princeton, New Jersey, 1957.

3. J. J. Florentin, "Optimal Control of Continuous Time, Markov Stochastic Systems," Jour. of Electronics and Control, Vol. 10, No. 6, p. 473, June 1961.

4. H. J. Kushner, "Optimal Stochastic Control", IRE Trans. on Automatic Control, Vol. AC-7, No. 5, October 1962.

5. E. Parzen, "Stochastic Processes", Holden Day Press, San Francisco, Calif., 1962.

6. C. W. Merriam III, "A Class of Optimum Control Systems," Jour. Franklin Institute, Vol. 267, 1959.

7. J. L. Bogdanoff and F. Kozin, "Moments of the Output of Linear Random Systems", Jour. Acoust. Soc. Amer., Vol. 34, No. 8, 1962, pp. 1063-1066.

8. M. S. Bartlett, "Stochastic Processes", Cambridge University Press, Cambridge, 1956, Sec. 3.5.

9. J. J. Florentin, "Optimal Control of Systems with Generalized Poisson Inputs", ASME Paper 62-JACC-1. Also in Jour. Basic Engr. (ASME) 1963.

A MODEL FOR FINITE-STATE PROBABILISTIC SYSTEMS

G. D. BRUCE
School of Electrical Engineering
Purdue University
Lafayette, Indiana, and
Poughkeepsie, New York

and

K. S. FU
Control and Information Systems Laboratory
School of Electrical Engineering
Purdue University
Lafayette, Indiana

Abstract

A mathematical model to describe finite-state probabilistic systems is presented. The description used is parallel to the conventional representation of finite-state deterministic systems. A probabilistic system can be represented by a table, a diagram or a matrix formulation. The usefulness of a matrix representation is that it permits reasonably flexible methods of computation. This type of description of a complex system facilitates the use of digital processing for handling an analysis by machine computation. Relationships between the proposed model and learning systems are discussed. The problem of estimating probabilistic parameters of a finite-state system from observation of its performance or its learning properties has been investigated. Probabilistic matrix techniques has been considered of value in analyzing system reliability and learning behavior, both of which are becoming increasingly important aspects of automata study.

Introduction

The development of the part of automata theory concerned with probabilistic systems has lagged behind the development of that for deterministic systems. The reason for this is that deterministic automata have been studied as the ideal, although no actual systems can ever reach this perfection. Furthermore, little has been presented in the form of mathematical problems that might be solved by probabilistic computation (other than simply random number generation) that cannot readily be done by a deterministic system. On the other hand, some important practical problems exist for the very reason that actual systems do not have zero failure probabilities, although these problems are not formulated in terms of system computability. Complications in systemsarise due to disturbances, component failure rates, imperfect maintenance, and other factors introducing a system failure probability that can be of great concern. In addition, the apparent probabilistic nature of biological processes seems to more than justify the investigation of statistically effective automata. Identification experiments, fault analysis, study of system learning process, and effectiveness of improvements in reliability in terms of successful runs of experiments, for example, may be of interest in probabilistic automata.

Mathematical Representation of Probabilistic Systems

The proposed representation of probabilistic systems will be restricted to the class of finite-state, fixed (not growing), discrete, and synchronous systems. The last restriction is a convenience so that a clocked system can be employed and the asynchronous problem of detecting the completion of one operation to start the next can be avoided [1].

Definition of a probabilistic automaton is given in much the same terms as that for a deterministic finite automaton [2], except for certain modifications to describe the probabilistic properties of the system.

Definition: A (finite) probabilistic automaton is a quintuple (I, V, Q, F, G) where

I is a non-empty finite set of objects called input states,

V is a non-empty finite set of objects called output states,

Q is a non-empty finite set of objects called internal states,

F is a probabilistic function (called the direct transition function) which associates with each $i \in I$ and $q \in Q$ and element $f(i, q) \in Q$, that is,

$$F : I \times Q \longrightarrow Q$$

and G is a probabilistic function (called the direct output function) such that

$$G : I \times Q \longrightarrow V$$

Let $I = \{ i_1, i_2, \ldots, i_p \}$

$V = \{ v_1, v_2, \ldots, v_r \}$

and $Q = \{ q_1, q_2, \ldots, q_n \}$

If q(k) and i(k) denote the internal state and the input to the automation, respectively, at time kt, then q(k + 1) and v(k + 1), the internal state and the output, respectively, at time (k + 1)t, will be given by

$$q(k + 1) = F[i(k), q(k)]$$
$$v(k) = G[i(k), q(k)]$$

A one-step internal state transition occurs for each input, i_h, of the input state set $\{ a, b, \ldots, h \}$. One-step transitions from any state q_i to any other state q_j may have a non-zero probability. The probabilities of transitions from q_i to q_j may be specified by a set of transition matrices, $\{[A], [B], \ldots, [H]\}$, depending on the input, i_h.

$$
\begin{aligned}
[A] &= [a_{ij}] \quad \text{where} \quad a_{ij} = p\{q_i(k) \rightarrow q_j(k+1) \mid a\} \\
[B] &= [b_{ij}] \quad \text{where} \quad b_{ij} = p\{q_i(k) \rightarrow q_j(k+1) \mid b\} \\
&\ \vdots \qquad\qquad\qquad\quad \vdots \\
[H] &= [h_{ij}] \quad \text{where} \quad h_{ij} = p\{q_i(k) \rightarrow q_j(k+1) \mid h\}
\end{aligned}
$$

$$i = 1, 2, \ldots, n \; ; \qquad j = 1, 2, \ldots, n.$$

The probability of a transition matrix being a member of the set $\{[A], [B], \ldots, [H]\}$ is unity. This follows that the probability of i_h being in $\{a, b, \ldots, h\}$ is unity. Furthermore, the probability of a transition from $q_i(k)$ to $\{q_1(k+1), \ldots, q_n(k+1)\}$ is unity. Therefore, for given $q_i(k)$

$$\sum_{j=1}^{n} a_{ij} = 1 \quad , \quad \sum_{j=1}^{n} b_{ij} = 1 \quad , \ldots , \quad \sum_{j=1}^{n} h_{ij} = 1$$

Since a square matrix with non-negative terms and unit row sums is called a "stochastic matrix", each matrix in the set $\{[A], [B], \ldots, [H]\}$ is a stochastic matrix. It is noted that if in each row of each of the matrices $\{[A], [B], \ldots, [H]\}$ there is one and only one non-zero term, then this term is unity. In this case, the sum of these matrices merely reduces to the transition matrix of a deterministic automaton in "skeleton matrix" form, provided each non-zero sum term is set equal to 1. Given an automaton D where for each matrix of $\{[A], [B], \ldots, [H]\}$,

$$h_{ij} = 1 \quad \text{for one } j \text{ and zero for all others in each row}$$
$$i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, n$$

Then, the skeleton matrix of D,

$$[\tilde{D}] = [A] + [B] + \ldots + [H]$$

where $\tilde{d}_{ij} = 1$ if non-zero, for a deterministic automaton. The deterministic transition table is constructed by entering the unity probability $q_j(k+1)$ from each of $\{[A], [B], \ldots, [H]\}$ for each $q_i(k)$ as the $(k+1)$th internal state following $q_i(k)$ for inputs $\{a, b, \ldots, h\}$. A similar procedure is also followed in describing the "most probable" operational automaton from a specified probabilistic automaton.

The probabilistic description of the output states, v_h, also requires a set of output probability matrices $\{[A_o], [B_o], \ldots, [H_o]\}$ corresponding to the inputs $i_h = \{a, b, \ldots, h\}$ respectively. The output states can be taken as dependent only on $q_i(k)$ and i_h, or else a dependence on $q_j(k + 1)$ can also be taken. The output set $\{[A_o], [B_o], \ldots, [H_o]\}$ does not affect the internal states and does not enter into more than a 1-step transition chain.

Fig. 1a, 2a, and 3a show the transition table, state diagram and "π-matrix" representation for describing deterministic automata. Parallel representations for describing probabilistic automata are illustrated in Fig. 1b, 2b, and 3b.

	$q(k + 1)$				$v_h(k)$			
$i_h(k)$ \ $q(k)$	a	b	c	. . .	a	b	c	. . .
q_1	$q(k + 1) = f[i(k), q(k)]$				$v(k) = g[i(k), q(k)]$			
q_2								
⋮								
q_n								

Fig. 1a. Deterministic Transition Table

	$q(k + 1)$				$v_h(k)$			
$i_h(k)$ \ $q(k)$	a	b	c	...	a	b	c	...
q_1	[A]	[B]	[C]	...	$[A_o]$	$[B_o]$	$[C_o]$	...
q_2								
⋮								
q_n								

Fig. 1b. Probabilistic Transition Table

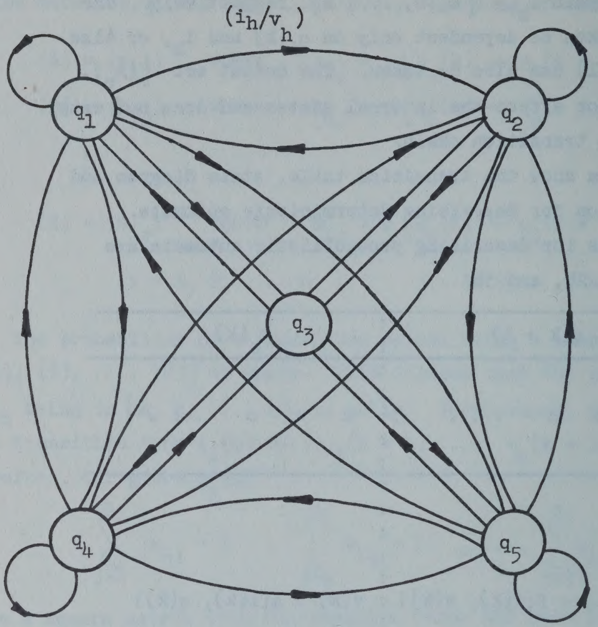


Fig. 2. Transition Diagram (n = 5)

2(a). Deterministic Case:

i) Each branch in the diagram is assigned a probability $P\{q_i \rightarrow q_j \mid i_h/v_h\}$ of either 1 or 0.
ii) If and only if $P\{\ \} = 1$ is the branch drawn, otherwise it is let off.
iii) The branches with $P\{\ \} = 1$ are then labeled with only inputs which produce the indicated transition and the resulting output, for example,

$$(i_1/v_1)v \dots v(i_h/v_h)$$

2(b). Probabilistic Case:

Each possible branch is shown and labeled, for example, $(P\{a_{ij} \mid a\}/v_{a_{ij}})v\ (P\{b_{ij} \mid b\}/v_{b_{ij}})v \dots v(P\{h_{ij} \mid h\}/v_{h_{ij}})$ with ij branch terms taken from the matrices [A], [B], ..., [H]. A simplified labeling would give only the value of the probability and the associated output with ordering according to the input table order $\{a, b, \dots, h\}$, for example, (0.99/1), (0.01/1) ... (0.90/1).

$$[P]^{(1)} = \begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1n} \\ \pi_{21} & \pi_{22} & & \cdot \\ \vdots & & & \cdot \\ & & & \cdot \\ \pi_{n1} & \cdot \; \cdot \; \cdot \; \cdot & \cdot \; \cdot & \pi_{nn} \end{bmatrix}$$

Fig. 3. "π - Matrix "

3(a). Deterministic Case:

$$\text{where } \pi_{ij} = \begin{cases} \text{branch label, if any} \\ \\ 0, \text{ if no branch} \end{cases}$$

$$[P]^{(1)} = [\text{1-step paths}]$$

$$[P]^{(k)} = \left\{ [P]^{(1)} \right\}^{k} = [P]^{(1)} \; [P]^{(k-1)} = [\text{k-step paths}]$$

If only proper paths are included, then

$$[P]^{(k)} = 0 \qquad \text{for } k < n - 1$$

3(b). Probabilistic Case:

The "π-matrix" is obtained merely by taking the most probable paths equal to unity probability, giving the equivalent of the deterministic π-matrix. The kth-power ordered matrix product gives the kth-step paths.

Input Sequence Transition Probability Matrix

If a probabilistic transition matrix is formed at each step of a sequence with the transition probabilities corresponding to the input of each step, the ordered product of these matrices yields the transition probabilities of the entire sequence. For example, the transition matrix after the 5-step sequence a a b c a has been applied to an automaton is

$$[T]^{(5)} = [A] \; [A] \; [B] \; [C] \; [A] = \prod_{i_h = i_1}^{i_5} [h_{ij}]$$

where $i_1, \ldots, i_5$ = a a b c a. Then, the initial probability matrix of a given state p_i, which is a diagonal matrix $[p_i \; \delta_{ij}]$, times $[T]^{(5)}$ gives the probability of the 5th step internal state $q_j(5)$. δ_{ij} denotes Kronecker delta.

$$[p_i\ \delta_{ij}][T]^{(5)} = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & & \vdots \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & p_n \end{bmatrix} \begin{bmatrix} t_{11}^{(5)} & t_{12}^{(5)} & \cdots & t_{1n}^{(5)} \\ t_{21}^{(5)} & & & \vdots \\ \vdots & & & \vdots \\ t_{n1}^{(5)} & \cdots & \cdots & t_{nn}^{(5)} \end{bmatrix}$$

$$= \begin{bmatrix} p_1\ t_{11}^{(5)} & p_1\ t_{12}^{(5)} & \cdots & p_1\ t_{1n}^{(5)} \\ p_2\ t_{21}^{(5)} & & & \vdots \\ \vdots & & & \vdots \\ p_n\ t_{n1}^{(5)} & \cdots & \cdots & p_n\ t_{nn}^{(5)} \end{bmatrix}$$

It is noted that this result is very similar to that which has been obtained from an ergodic Markov chain [3, 4]. In the ergodic Markov case, an initial transition probability matrix [P] is given. The matrix [P] does not change, however, for a sequence of steps in a Markov chain. Furthermore, a stationary matrix [S] can be defined such that statistical equilibrium is reached when

$$\lim_{k \to \infty} [P]^k = [S], \quad \text{with equal column terms.}$$

To solve for the elements, s_j, of [S], we can write [S] = [S][P] since

$$\lim_{k \to \infty} [P]^{k-1} \cdot [P] = [S][P] .$$

This leads to the set of equations

$$s_j = \sum_{\alpha=1}^{n} s_\alpha\ p_{\alpha j}\ , \quad (j = 1, \ldots, n)$$

which can be solved for $s_1, s_2, \ldots, s_n$ for given $[p_{ij}]$.

It has been shown that, associated with an input sequence a a b c a, there is a transition matrix $[T]^{(5)}$. If the input sequence is repeated, a stationary matrix is approached. In general, a repeated sequence of length r has a transition matrix $[T]^{(r)}$. If this sequence is repeated m times, the transition matrix is $[[T]^{(r)}]^{(m)}$. Furthermore,

$$\lim_{m \to \infty} [[T]^{(r)}]^{(m)} = [S]^{(r)}$$

which is a stationary transition probability matrix. The elements, $s_j^{\,r}$, of $[S]^{(r)}$ can obviously be solved for, as above for an ergodic Markov chain. An important effect which is emphasized by this is the eventual approach to equal column terms in the transition matrix, or the loss of almost all information about the initial states, if the process continues through sufficient steps.

If the input is also probabilistic, the representation of an automaton can be reduced to a single transition probability matrix. For each possible input $\{a, b, \ldots, h\}$ a probability is assigned $\{P(a), P(b), \ldots, P(h)\}$, where $P(a) + P(b) + \ldots + P(h) = 1$. These probabilities time their respective transition matrices $\{[A], [B], \ldots, [H]\}$, resulting in the new matrices $\{[P(a)\ a_{ij}], [P(b)\ b_{ij}], \ldots, [P(h)\ h_{ij}]\}$. By taking the sum of these resulting matrices, an over-all transition probability matrix of the automaton is obtained. For example, given internal states of an automaton $\{q_1, \ldots, q_5\}$, and input states $\{a, \ldots, d\}$, with $P(a) = p_a, \ldots, P(d) = p_d$, the overall probability matrix is

$$S^{(1)} = \left[p_a[A] + p_b[B] + \ldots + p_d[D] \right]$$

$$= \begin{bmatrix} (p_a\ a_{11} + \ldots + p_d\ d_{11}) & \ldots & (p_a\ a_{15} + \ldots + p_d\ d_{15}) \\ \vdots & \vdots & \vdots \\ (p_a\ a_{51} + \ldots + p_d\ d_{51}) & \ldots & (p_a\ a_{55} + \ldots + p_d\ d_{55}) \end{bmatrix}$$

The k-step transition probability matrix becomes

$$S^{(k)} = S^{(1)} \cdot S^{(k-1)} = \left[S^{(1)} \right]^k$$

This shows that an automaton with probabilistic transition and probabilistic input can be described by a single probability matrix with the sequencing being represented by a Markov process. The significance of a probability distribution for the inputs is that it facilitates an investigation of the internal state probabilities for different input sequence lengths. More generally, however, interest would seem to be concentrated on questions of what probable internal states will occur for given input sequences and not for probabilistic inputs.

System Reliability and Identification Experiment

One objective of the investigation of probabilistic automata is to analyze the reliability of an automaton and, consequently, to design a reliable automaton from unreliable components. In this case, the probability of correct functioning of each component, p, must be reasonably high, even though it is not unity. This means that the "intended" function of an automaton can be determined from inspecting the representation for high-probability transition paths. Based on this concept the structure of the "intended automaton" can be found if it is not already known. When this is done, minimization procedures can be followed to obtain the minimal form of the intended automaton. Next, a new probabilistic representation for this minimal automaton can be made on the basis of the reliability distributions of components.

A simple example of a probabilistic experiment is given in Fig. 4. Fig. 4a shows an "intended" automaton which would be derived from the total probabilistic automaton shown in Fig. 4b, in state diagram form. Fig. 4c illustrates the transition probability matrices with the transition table structure implied. The two-step transition probability matrices are shown in Fig. 5 for inputs a a and a b. The decrease in the probability of a successful run for increasing experiment length is already apparent. Also, for three different "parts" reliabilities the two results, a a and a b, can be compared for, say, the probability of one sequence giving the same result as the other one.

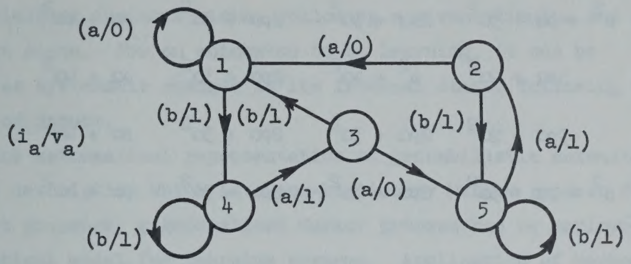
State identification experiments can be performed in terms of the "intended automaton" model. The probabilistic representation is employed to determine the probability of the automaton working as intended. The results of the experiment can be used with established confidence levels for both the initial state and the final state identification problems. The automaton identification procedures of an "intended" (probabilistic) automaton follow in parallel with the procedures for the identification of a deterministic automaton. The interest in this lies in the probability of a successful identification. Most of the identification theorems for deterministic automata can also be applied to an "intended" probabilistic automaton. Closely allied to the automaton identification problem is the fault diagnosis problem. The fault diagnosis problem is an identification problem which reduces to a final-state identification experiment. The term "faulty automaton" means a permanent fault has occurred which, in effect, causes the automaton to act like another automaton of a known class.

Learning Process in a Probabilistic System

Learning can be considered as "orderly changes in response probabilities resulting from the occurrence of events" [5]. From the

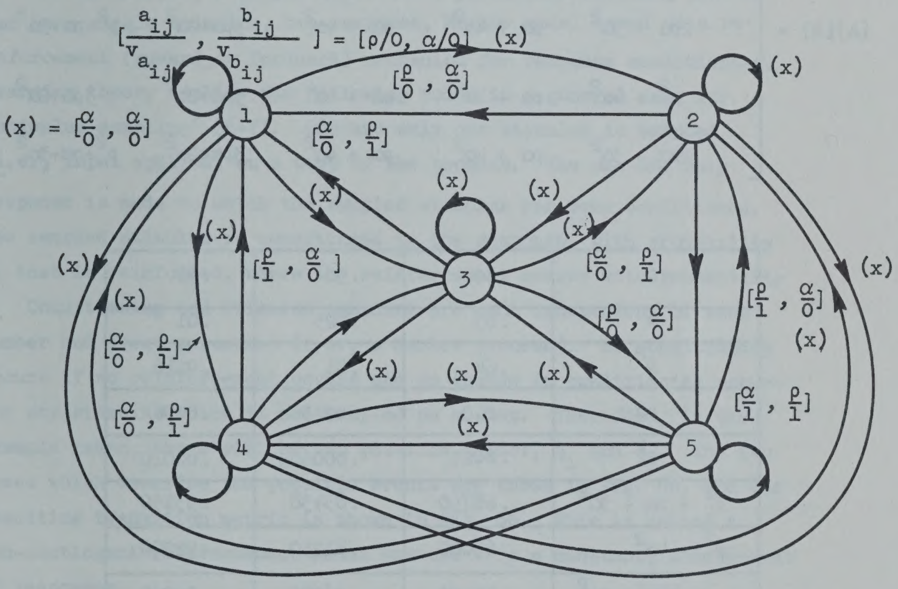
Figure 4

(4a) Deterministic ("Intended" Automaton)


(a/0)
(a/0)
1
2
(b/1)
(b/1)
3
(b/1)
(a/1)
(i_a/v_a)
4
(a/1)
(a/0)
5
(b/1)
(b/1)

(4b) Probabilistic Representation

State Diagram: Let $\rho < 1$, and $\alpha = \frac{1 - \rho}{4} > 0$


$[\frac{a_{ij}}{v_{a_{ij}}}, \frac{b_{ij}}{v_{b_{ij}}}] = [\rho/0, \alpha/0]$ (x)
(x)
$[\frac{\rho}{0}, \frac{\alpha}{0}]$
1
2
$(x) = [\frac{\alpha}{0}, \frac{\alpha}{0}]$
$[\frac{\alpha}{0}, \frac{\rho}{1}]$
(x)
(x)
(x)
(x)
(x)
$[\frac{\rho}{1}, \frac{\alpha}{0}]$
3
$[\frac{\alpha}{0}, \frac{\rho}{1}]$
(x)
(x)
$[\frac{\rho}{1}, \frac{\alpha}{0}]$
(x)
$[\frac{\alpha}{0}, \frac{\rho}{1}]$
$[\frac{\rho}{0}, \frac{\alpha}{0}]$
(x)
(x)
(x)
(x)
$[\frac{\alpha}{0}, \frac{\rho}{1}]$
$[\frac{\alpha}{1}, \frac{\rho}{1}]$
(x)
4
5

(4c) Matrices:

$$A = \begin{bmatrix} \rho & \alpha & \alpha & \alpha & \alpha \\ \rho & \alpha & \alpha & \alpha & \alpha \\ \alpha & \alpha & \alpha & \alpha & \rho \\ \alpha & \alpha & \rho & \alpha & \alpha \\ \alpha & \rho & \alpha & \alpha & \alpha \end{bmatrix}, \quad B = \begin{bmatrix} \alpha & \alpha & \alpha & \rho & \alpha \\ \alpha & \alpha & \alpha & \alpha & \rho \\ \rho & \alpha & \alpha & \alpha & \alpha \\ \alpha & \alpha & \alpha & \rho & \alpha \\ \alpha & \alpha & \alpha & \alpha & \rho \end{bmatrix}, \quad A_o = \begin{bmatrix} 4\alpha & \rho_1 \\ 4\alpha & \rho_1 \\ 4\alpha & \rho_5 \\ \rho_3 & 4\alpha \\ \rho_2 & 4\alpha \end{bmatrix}, \quad B_o = \begin{bmatrix} \rho_4 & 4\alpha \\ \rho_5 & 4\alpha \\ \rho_1 & 4\alpha \\ \rho_4 & 4\alpha \\ \rho_5 & 4\alpha \end{bmatrix}$$

Note: Subscripts indicate $q(k + 1)$

$$[A][A] = \begin{bmatrix} \rho^2 + \rho\alpha + 3\alpha^2 & 2\rho\alpha + 3\alpha^2 & 2\rho\alpha + 3\alpha^2 & \rho\alpha + 4\alpha^2 & 2\rho\alpha + 3\alpha^2 \\ \rho^2 + \rho\alpha + 3\alpha^2 & 2\rho\alpha + 3\alpha^2 & 2\rho\alpha + 3\alpha^2 & \rho\alpha + 4\alpha^2 & 2\rho\alpha + 3\alpha^2 \\ 3\rho\alpha + 2\alpha^2 & \rho^2 + 4\alpha^2 & 2\rho\alpha + 3\alpha^2 & \rho\alpha + 4\alpha^2 & 2\rho\alpha + 3\alpha^2 \\ 3\rho\alpha + 2\alpha^2 & 2\rho\alpha + 3\alpha^2 & 2\rho\alpha + 3\alpha^2 & \rho\alpha + 4\alpha^2 & \rho^2 + 4\alpha^2 \\ \rho^2 + \rho\alpha + 3\alpha^2 & 2\rho\alpha + 3\alpha^2 & 2\rho\alpha + 3\alpha^2 & \rho\alpha + 4\alpha^2 & 2\rho\alpha + 3\alpha^2 \end{bmatrix}$$

$$[A][B] = \begin{bmatrix} 2\rho\alpha + 3\alpha^2 & \rho\alpha + 4\alpha^2 & \rho\alpha + 4\alpha^2 & \rho^2+\rho\alpha+3\alpha^2 & 2\rho\alpha+2\alpha^2 \\ 2\rho\alpha + 3\alpha^2 & \rho\alpha + 4\alpha^2 & \rho\alpha + 4\alpha^2 & \rho^2+\rho\alpha+3\alpha^2 & 3\rho\alpha+2\alpha^2 \\ 2\rho\alpha + 3\alpha^2 & \rho\alpha + 4\alpha^2 & \rho\alpha + 4\alpha^2 & 3\rho\alpha+2\alpha^2 & \rho^2+\rho\alpha+3\alpha^2 \\ \rho^2 + 4\alpha^2 & \rho\alpha + 4\alpha^2 & \rho\alpha + 4\alpha^2 & 3\rho\alpha+2\alpha^2 & 3\rho\alpha+2\alpha^2 \\ 2\rho\alpha + 3\alpha^2 & \rho\alpha + 4\alpha^2 & \rho\alpha + 4\alpha^2 & 3\rho\alpha+2\alpha^2 & \rho^2+\rho\alpha+3\alpha^2 \end{bmatrix}$$

ρ	.80	.90	.96
α	.05	.025	.01
ρ^2	.64	.81	.9216
$\rho\alpha$	.040	.0225	.0096
α^2	.0025	.000625	.00010
$\rho^2 + \rho\alpha + 3\alpha^2$	.68750	.83438	.93150
$\rho^2 + 4\alpha^2$	.65000	.81250	.92200
$3\rho\alpha + 2\alpha^2$	.12500	.06975	.02900
$2\rho\alpha + 3\alpha^2$	.08750	.04748	.01950
$\rho\alpha + 4\alpha^2$	.05000	.02500	.01000

Fig. 5. Two-Step Transition Probability Matrices - Example

probabilistic representation of a system developed above, it is easy to see the relation of the system model to a learning process. The inputs are the stimuli. The internal states can be considered as the response probabilities. The outputs are the observed responses or reactions. The transition probabilities are the event probabilities of an outcome yielding the next state, following a given stimulus and a given present state. For an automaton to be learning, it can be interpreted as systematic changes in its internal states following the application of inputs.

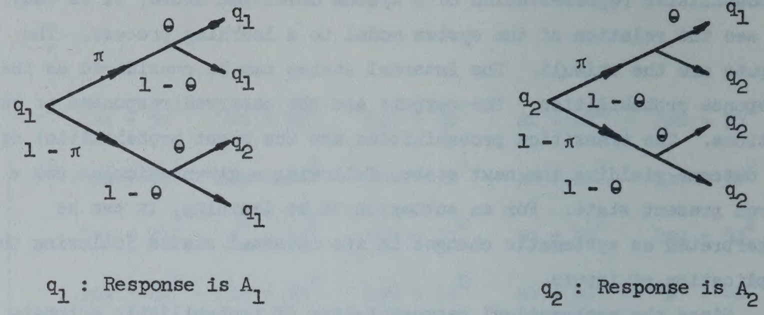
Since the mathematical representation of probabilistic automata results in a description which is a non-stationary Markov process for a given input sequence, a generalized Markov process can be employed as a mathematical model for learning systems. Application of Markov processes to learning theory has been treated in the literature [5-7]. The approach can be directly extended to include the learning automata.

As an example, a simple model for a particular learning process can be a single stimulus, two-response, Markov model based on a reinforcement (reward or feedback) mechanism for response conditioning. Learning theory applies the following rules in a general case for "stimulus sampling" [6-7]. One and only one stimulus is sampled (i.e., input applied) on a step of the process. The one and only response is made to which the sampled stimulus has been conditioned. The sampled stimulus is conditioned to the response, with probability θ, that is reinforced, where the reinforcement occurs with probability π. Conditioning and stimulus sampling are each independent of step number and event sequences (i.e., a Markov process). No conditioning occurs if no reinforcement occurs and no change in conditioning occurs for any stimulus which is not sampled on a step. Therefore, in the example taken, there will be two possible states, q_1 and q_2. The two trees which describe the possible events are shown in Fig. 6a, and the resulting transition matrix is shown in Fig. 6b. This is called a non-contingent reinforcement model because π is a constant, independent of responses.

Next, a state diagram for the system is shown in Fig. 6c and the transition matrix for a learning automaton matching this model is given in Fig. 7a. The transition matrix is stochastic and stationary, or asymptotic, response probabilities are π and $1 - \pi$. A typical mean learning curve is shown in Fig. 7b.

Statistical Estimation of Parameters

Statistical analysis of probabilistic automata can be made following known procedures for analyzing Markov processes [8]. Statistical a)


θ
q_1
π
1 - θ
q_1
q_1
1 - π
θ
q_2
1 - θ
q_1
θ
q_1
π
q_2
q_2
1 - θ
1 - π
θ
q_2
1 - θ
q_2
q_1 : Response is A_1
q_2 : Response is A_2

b)

$$\begin{array}{c} \\ q_1 \\ q_2 \end{array}\begin{array}{c} \begin{array}{cc} q_1 & q_2 \end{array} \\ \left[\begin{array}{cc} \theta\pi + (1 - \theta) & \theta(1 - \pi) \\ \theta\pi & 1 - \theta\pi \end{array}\right] \end{array} \quad \text{(Stimulus: } S_1\text{)}$$

c)

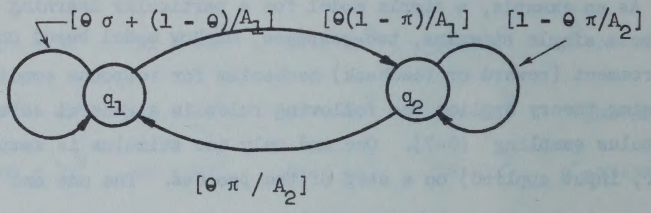

$[\theta\sigma + (1 - \theta)/A_1]$
$[\theta(1 - \pi)/A_1]$
$[1 - \theta\pi/A_2]$
q_1
q_2
$[\theta\pi / A_2]$

Fig. 6. Learning Theory and System Model

tests of interest would be for transition probabilities, stationarity and order of a Markov process.

A set of a statistically identical runs of length (r + 1) will yield data on the states in the form:

$$q_o^{(1)}, q_1^{(1)}, \ldots, q_r^{(1)}$$
$$q_o^{(2)}, q_1^{(2)}, \ldots, q_r^{(2)}$$
$$\vdots$$
$$q_o^{(n)}, q_1^{(n)}, \ldots, q_r^{(n)}$$

where the superscript denotes the run and the subscript denotes the observation step number. Observations may be considered at times

		Input = S_1		Outputs	
	(k)	(k + 1) q_1	q_2	(k) A_1	A_2
a) Internal States	q_1	$\theta \pi + (1 - \theta)$	$\theta(1 - \pi)$	1	0
	q_2	$\theta \pi$	$1 - \theta \pi$	0	1

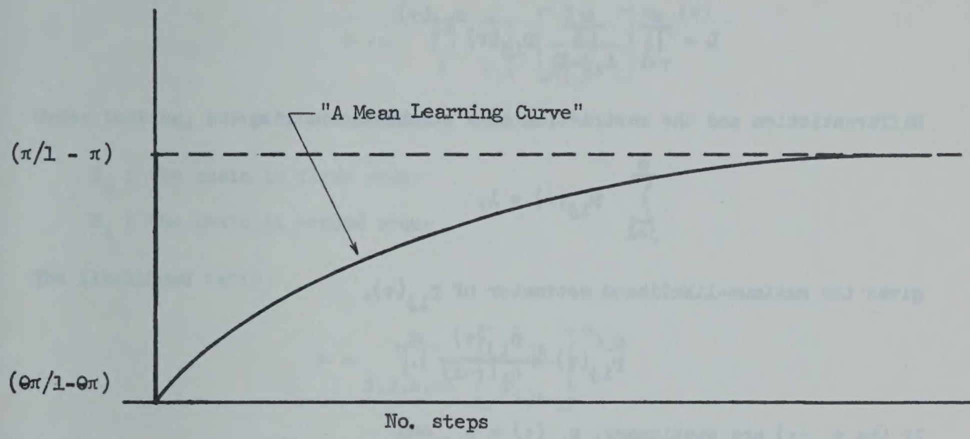


Fig. 7. Learning Automata

b) $$\frac{P_r[q(k+1) = q_1 \mid q(1) = q_2]}{P_r[q(k+1) = q_2 \mid q(1) = q_2]}$$

$\tau = 0, 1t, 2t, \ldots, rt$, as the times at which inputs are applied. (Each "observation" may, in turn, require an additional estimation in order to record each state as one in an admissible set of possible response classes, but the states are considered directly observable.)

Let $n_{ij}(\tau)$ be the number of runs in state i at time $(\tau - 1)$ and in state j at time τ. Let $n_i(\tau - 1)$ be the number of runs in state i at time $(\tau - 1)$. Then,

$$n_i(\tau - 1) = \sum_{b=1}^{m} n_{ib}(\tau) \quad ,$$

where the discrete states are 1, 2, ..., m.

Given $n_i(\tau - 1)$, the probability of observing $n_{i1}(\tau)$ transitions to state 1, $n_{i2}(\tau)$ transitions to state 2, ..., $n_{im}(\tau)$ transitions to state m, can be given by the multinomial formula

$$\frac{n_i(\tau - 1)\,!}{\prod_{j=1}^{m} n_{ij}(\tau)\,!} = \prod_{j=1}^{m} p_{ij}(\tau)^{n_{ij}(\tau)}$$

where p_{ij} is the probability of transition from state i to state j. The likelihood function, L, for the given data can be written

$$L = \prod_{\tau=1}^{r} \prod_{i,j=1}^{m} p_{ij}(\tau)^{n_{ij}(\tau)} \quad .$$

Differentiation and the restriction of a stochastic matrix,

$$\sum_{j=1}^{m} p_{ij}(\tau) = 1,$$

gives the maximum-likelihood estimator of $p_{ij}(\tau)$,

$$\hat{p}_{ij}(\tau) = \frac{n_{ij}(\tau)}{n_i(\tau-1)} \quad .$$

If the $p_{ij}(\tau)$ are stationary, $p_{ij}(\tau) = p_{ij}$ and

$$\hat{p}_{ij} = \frac{\sum_{\tau=1}^{r} n_{ij}(\tau)}{\sum_{\tau=0}^{r-1} n_i(\tau)} \quad .$$

Similarly, if the chain is second-order,

$$P_{ijk} = \frac{n_{ijk}}{\sum_{f=1}^{m} n_{ijf}} = \frac{\sum_{\tau=2}^{r} n_{ijk}(\tau)}{\sum_{\tau=1}^{r-1} n_{ij}(\tau)}$$

Stationarity tests are based on the above estimators. Assuming a first order Markov chain, the hypothesis testing is:

$$H_o : p_{ij}(\tau) = p_{ij} \qquad \tau = 1, 2, \ldots, r.$$

$$H_1 : p_{ij}(\tau) \text{ is dependent on } \tau.$$

Under H_o, $L = \prod_{\tau} \prod_{i,j} p_{ij}^{\;n_{ij}(\tau)}$

Under H_1, $L = \prod_{\tau} \prod_{i,j} p_{ij}(\tau)^{n_{ij}(\tau)}$.

The likelihood ratio,

$$\bar{\lambda} = \prod_{\tau} \prod_{i,j} \left[\frac{\hat{p}_{ij}}{\hat{p}_{ij}(\tau)} \right]^{n_{ij}(\tau)}$$

Order testing, assuming stationarity, tests:

H_o : the chain is first order

H_1 : the chain is second order

The likelihood ratio,

$$\lambda = \prod_{i,j,k,=1}^{m} \left[\frac{\hat{p}_{jk}}{\hat{p}_{ijk}} \right]^{n_{ijk}}$$

In the above, - 2 log $\bar{\lambda}$ is asymptotically distributed as x^2 with $(r - 1)[m(m - 1)]$ degrees of freedom for H_o true, and - 2 log λ is asymptotically distributed as x^2 with $m(m - 1)^2$ degrees of freedom for H_o true.

For an example, data will be used that resulted from tests in the field of social science on sequential changes in attitudes. The definition has been given, "An attitude is considered to be simply a choice of a specified set of responses to a question." Thus, data on attitudes directly corresponds to that on observed responses to stimuli (the questions) in our general learning system.

The data gives the responses at six discrete times, $\tau = 0$ thru $\tau = 5$, of each of 445 separate subjects. Changes in response probabilities, i.e., state transitions, are controlled by external factors which influence the opinions and give the transition probabilities from one state, specific opinion, or response class, to the next state.

Three response classes are taken, which might be "yes", "no", and ("maybe, indecision, or no opinion") for instance. These correspond to the states 1, ..., m where m = 3.

The given data and some results follow.

$n_i(\tau - 1)$	$n_{ij}(\tau)$			
$n_1(0) = 146$	125	5	16	$\tau = 1$
$n_2(0) = 128$	7	106	15	
$n_3(0) = 171$	11	18	142	
$n_1(1) = 143$	124	3	16	$\tau = 2$
$n_2(1) = 129$	6	109	14	
$n_3(1) = 173$	22	9	142	
$n_1(2) = 152$	146	2	4	$\tau = 3$
$n_2(2) = 121$	6	111	4	
$n_3(2) = 172$	40	36	96	
$n_1(3) = 192$	184	1	7	$\tau = 4$
$n_2(3) = 149$	4	140	5	
$n_3(3) = 104$	10	12	82	
$n_1(4) = 198$	192	1	5	$\tau = 5$
$n_2(4) = 153$	2	146	5	
$n_3(4) = 94$	11	12	71	

$n_1(5) = 205$
$n_2(5) = 159$
$n_3(5) = 81$

where

$$n_i(\tau - 1) = \sum_{j=1}^{3} n_{ij}(\tau)$$

and

$$n_j(\tau) = \sum_{i=1}^{3} n_{ij}(\tau)$$

$$[\hat{P}](1) = [\hat{p}_{ij}(1)] = \begin{vmatrix} 0.856 & .034 & .110 \\ 0.055 & .828 & .117 \\ 0.064 & .105 & .831 \end{vmatrix}$$

$$[\hat{P}](2) = [\hat{p}_{ij}(2)] = \begin{vmatrix} .867 & .021 & .112 \\ .047 & .845 & .108 \\ .127 & .052 & .821 \end{vmatrix}$$

$$[\hat{P}](3) = [\hat{p}_{ij}(3)] = \begin{vmatrix} .961 & .013 & .026 \\ .050 & .917 & .033 \\ .233 & .209 & .558 \end{vmatrix}$$

$$[\hat{P}](4) = [\hat{p}_{ij}(4)] = \begin{vmatrix} .958 & .005 & .037 \\ .027 & .940 & .033 \\ .096 & .115 & .789 \end{vmatrix}$$

$$[\hat{P}](5) = [\hat{p}_{ij}(5)] = \begin{vmatrix} .970 & .005 & .025 \\ .013 & .954 & .033 \\ .117 & .128 & .755 \end{vmatrix}$$

where

$$\hat{p}_{ij}(\tau) = \frac{n_{ij}(\tau)}{n_i(\tau - 1)}$$

If stationarity is assumed,

$$\hat{p}_{ij} = \frac{\sum\limits_{\tau=1}^{5} n_{ij}(\tau)}{\sum\limits_{\tau=0}^{4} n_i(\tau)} ,$$

$$[\hat{P}] = [\hat{p}_{ij}] = \begin{vmatrix} 0.928 & .014 & .058 \\ 0.037 & .900 & .063 \\ 0.132 & .122 & .746 \end{vmatrix}$$

Calculating

$$\bar{\lambda} = \prod_{\tau} \prod_{ij} \left[\frac{\hat{p}_{ij}}{\hat{p}_{ij}(\tau)} \right]^{n_{ij}(\tau)}$$

$$\bar{\lambda} = \prod_{\tau} \prod_{i} \prod_{j} \left[\frac{\sum\limits_{\tau=1}^{5} n_{ij}(\tau) \Big/ \sum\limits_{\tau=0}^{4} n_i(\tau)}{n_{ij}(\tau) \,/\, n_i(\tau - 1)} \right]^{n_{ij}(\tau)}$$

Evaluation of this is tedious if not handled by machine computation. In this case, there is a product of 3 x 3 x 5 = 45 factors each of which is a quotient raised to a power of from one to nearly two hundred. In logarithmic form,

$$\log \bar{\lambda} = \sum_{\tau=1}^{5} \sum_{i=1}^{3} \sum_{j=1}^{3} n_{ij}(\tau) \log [\hat{p}_{ij} \,/\, \hat{p}_{ij}(\tau)]$$

This results in

$$-2 \log \bar{\lambda} = 101.512 \ .$$

By taking $(r-1) = 4$, $m = 3$, and $(m-1) = 2$, we look under $(r-1)(m)(m-1) = 24$ degrees of freedom in a χ^2 distribution table. The probability of a value of 101.512 is much less than .05 and the H_0 is rejected.

This result might have been anticipated since only the first few steps of the process were taken, that is, no asymptotic approach to a final matrix was achieved. Furthermore, the nature of the process was such that a variation in the step by step operator (non-linear) might be expected. Otherwise, if H_0 were accepted, a stationary $[\hat{P}]$ would result in asymptotically approaching a set of states with probabilities of $\hat{q}_i$ according to a column vector $[\hat{Q}]$, where

$$[\hat{Q}][\hat{P}] = [\hat{Q}] \ .$$

The $\hat{q}_i$ can be found by solving the equations

$$\sum_{\alpha=1}^{3} \hat{q}_\alpha \hat{p}_{\alpha j} = \hat{q}_i \quad , \quad (i = j = 1, 2, 3)$$

For the $[\hat{P}]$ obtained, this gives

$$[\hat{Q}] = \begin{bmatrix} .506 \\ .303 \\ .191 \end{bmatrix} \ .$$

Conclusion

A finite-state probabilistic system can be represented by a matrix formulation, transition table or state diagram. Each input state of a system has an associated internal state transition probability matrix and a corresponding output matrix. Distinct transition matrices, which, in general, are different for each input state of a system, result in a description of system operation which is a non-stationary Markov process.

This mathematical description of a probabilistic system as a generalized Markov chain leads to a model for learning automata. Learning theory has employed Markov processes to describe the behavior of subjects during tests for learning, and each stimulus for a learning subject corresponds to an input state for a learning system. Similarly, each response probability corresponds to an internal state and observed responses correspond to output state of the system.

Experiments on probabilistic systems pose the problem of parameter determination. Statistical estimation procedures for Markov chain parameters can be effectively employed. Statistical tests for stationarity, transition probabilities, asymptotic probabilities and order of the process can be made.

The mathematical representations for probabilistic automata have been related to mathematical models employed in the behavioral sciences. The resulting techniques for system analysis are of interest in obtaining improved analytical methods in the increasingly important areas of reliability, prediction, and the learning properties of probabilistic systems.

References

1. "Analysis of Sequential Machines," D. D. Aufenkamp and F. E. Hohn, PGEC, December, 1957, Vol. EC-6, No. 4.

2. "An Introduction to Mathematical Machine Theory," Seymour Ginsburgh, Addison_Wesley, 1962.

3. "An Introduction to Probability Theory and its Applications," W. Feller, Vol. I, Chaps. XV and XVI, 2nd. Ed., Wiley, 1957.

4. "An Introduction to Information Theory," F. M. Reza, McGraw-Hill, 1961.

5. "Stochastic Models for Learning," R. R. Bush and F. Mosteller, John Wiley, 1955.

6. "Studies in Mathematical Learning Theory," R. R. Bush and W. K. Estes, Stanford Press, 1959.

7. "Markov Learning Models for Multiperson Interactions," and Atkinson, Stanford Press, 1960.

8. "Statistical Inference about Markov Chains," T. W. Anderson and L. A. Goodman, Annals of Math. Stat., March, 1957, Vol. 28, No. 1.

9. "Mathematical Thinking in the Social Sciences," edited by P. F. Lazarsfeld, The Free Press, Glencoe, Illinois, 2nd. Ed., 1955 (Chap. I, "Probability Models for Analyzing Time Changes in Attitudes," T. W, Anderson).

ORTHONORMAL FUNCTIONS IN ADAPTIVE DISCRETE FUNCTIONS

D. H. S. HSU
New York Transit Authority
New York

and

M. D. SRINATH and Y. FU
The University of Kansas
Lawrence, Kansas

Adaptive control is sometimes defined as control involving automatic measurement of the signals or process dynamics and the subsequent automatic self-design of the control system. A control system can be adequately designed only if the dynamic characteristics of the process to be controlled are known. If these characteristics are not known, it becomes necessary to measure the inputs and responses to generate the dynamic characteristics of the process. For achieving an identification of the process, an expansion over a finite number of functions of an orthogonal set can be used.

For a completely discrete system, the values of the signals are of interest only at discrete intervals of time corresponding to the sampling time in the system. Thus an expansion over a set of weighted orthonormal discrete functions is needed. A set of continuous orthonormal functions can be constructed from a set of component functions by the use of the Schmidt procedure.[1] For a set of discrete functions $\left\{\phi_i(kT)\right\}$, the orthogonality condition with respect to a weighting function $w(t)$ is given by

$$\sum_{k=0}^{\infty} \phi_i(kT)\, \phi_j(kT)\, w(kT) = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases} \qquad (1)$$

Let $\phi_i(kT)\sqrt{w(kT)} = \theta_i(kT)$ (2)

Then, the orthogonality condition in the z-domain will be given as

$$\frac{1}{2\pi jT} \oint_{\Gamma} \theta_i^{c}\left(\frac{1}{z}\right)\, \theta_j(z)\, z^{-1}\, dz = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases} \qquad (3)$$

where the superscript c refers to the conjugate function.

As pointed out by Young and Huggins,[2] most of the functions which are orthogonal in the Laplace domain are unfortunately not orthogonal in the z-domain since the orthonormal property of a set of functions is not preserved under the sampling operation.

In order to eliminate the need for large memories in the system, w(t) must be chosen such that only a few terms in the expansion over the set are adequate to identify the process. Thus a suitable choice for the weighting function would be a decaying exponential $w(kT) = e^{-2akT}$. With this choice of the weighting function, the orthogonality condition in the z-domain becomes

$$\frac{1}{2\pi jT} \oint_{\Gamma} \phi_i\left(\frac{1}{ze^{-aT}}\right)^c \phi_j(ze^{-aT})\, z^{-1}\, dz = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases} \qquad (4)$$

The set of functions $\{\phi_i(z)\}$ can be constructed on the basis of a suitable choice of poles in the z-domain by using a procedure similar to that described by Young and Huggins. Thus the first two members of the set will be given by

$$\phi_1(z) = b_1 \frac{z^2 - z}{(ze^{-aT} - z_1)(ze^{-aT} - z_1{}^c)} \qquad (5)$$

and

$$\phi_2(z) = b_2 \frac{z^2 + z}{(ze^{-aT} - z_1)(ze^{-aT} - z_1{}^c)} \qquad (6)$$

where the normalizing constants b_1 and b_2 are given by

$$b_1 = \left\{ \frac{Te^{aT}}{2} (e^{-2aT} - |z_1|^2) \left[e^{-2aT} + (z_1 + z_1{}^c)e^{-aT} + |z_1|^2 \right] \right\}^{\frac{1}{2}} \qquad (7)$$

and

$$b_2 = \left\{ \frac{Te^{aT}}{2} (e^{-2aT} - |z_1|^2) \left[e^{-2aT} - (z_1 + z_1{}^c)e^{-aT} + |z_1|^2 \right] \right\}^{\frac{1}{2}} \qquad (8)$$

Using the concept of the complementary operator in the construction of orthonormal functions other than the first pair, a set of discrete orthonormal functions with more than one pair of poles may be written as [1,2]

$$\phi_{2k-1}(z) = b_{2k-1} \frac{z^2 - z}{(ze^{-aT} - z_k)(ze^{-aT} - z_k^c)} \prod_{i=1}^{k} z_i z_i{}^c \frac{\left(z - \frac{e^{-aT}}{z_i}\right)\left(z - \frac{e^{-aT}}{z_i{}^c}\right)}{(ze^{-aT} - z_i)(ze^{-aT} - z_i{}^c)}$$

and

$$\phi_{2k}(z) = b_{2k} \frac{z^2 + z}{(ze^{-aT}-z_k)(ze^{-aT}-z_k^c)} \prod_{i=1}^{k} z_i z_i^c \frac{(z - \frac{e^{-aT}}{z_i})(z - \frac{e^{-aT}}{z_i^c})}{(ze^{-aT}-z_i)(ze^{-aT}-z_i^c)} \quad (9)$$

where

$$b_{2k-1} = \left\{ \frac{Te^{aT}}{2} (e^{-2aT} - |z_k|^2) \left[e^{-2aT} + (z_k + z_k^c) e^{-aT} + |z_k|^2 \right] \right\}^{\frac{1}{2}}$$

$$b_{2k} = \left\{ \frac{Te^{aT}}{2} (e^{-2aT} - |z_k|^2) \left[e^{-2aT} - (z_k + z_k^c) e^{-aT} + |z_k|^2 \right] \right\}^{\frac{1}{2}} \quad (10)$$

The poles in the z-domain are related to the poles in the s-domain such that

$$z_k = e^{s_k T} = e^{(-\alpha_k + j\beta_k)T}$$

$$z_k^c = e^{s_k^c T} = e^{(-\alpha_k - j\beta_k)T} \quad (11)$$

As the sampling interval T approaches zero, the discrete orthonormal functions will approach the ordinary continuous orthonormal functions.

The Adaptive System

The orthonormal functions constructed above will now be used to describe an adaptive scheme which may be used to design a system. The system is designed to minimize a certain measure of the error, the error being the difference between a desired output and the actual output. If the system functions of the plant are not known or are known only approximately, it becomes necessary to adequately describe the process dynamics. Once this is done control loops have to be set up which minimize the measure of performance. It is assumed in the rest of the discussion that the system is completely discrete, so that signals need be considered only at sampling instants.

In order to obtain a mathematical model which describes process dynamics, an expansion of the system function in a finite number of terms of an infinite series is often useful.[3] Expansion in a set of functions which are orthonormal with respect to a weighting function yields a minimum mean-squared error. For continuous systems, expansion in a set of orthonormal functions with weighting function $w(t) = e^{-2aT}$ has been proposed in the literature.[1,3] Suppose the process dynamics of the system under consideration can be expressed in terms of an impulse response $g(kT)$. Then, $g(kT)$ can be expanded in terms of the orthonormal set $\{\phi_i(kT)\}$ as

$$g(kT) = \sum_{i=1}^{N} c_i\phi_i(kT) \tag{12}$$

and z-transformation of (12) gives

$$G(z) = \sum_{i=1}^{N} c_i\phi_i(z) \tag{13}$$

In order to identify a process in terms of a specified set of orthogonal functions, it is necessary to construct a spectrum analyzer which operates on the signal to be expanded in such a way as to obtain the spectrum (the coefficients c_i) of the signal.[3]

As shown in Figure 1, let the impulse response of the system be $g(t)$, then the relation between the output of the system, $h^*(t)$, and the input $m^*(t)$ is given by

$$h^*(t) = \sum_{k=0}^{\infty} m(kT)\ g(t-kT) \tag{14}$$

From (13) it can be seen that the signal input to a system can also be described in terms of the orthonormal functions $\phi_i(kT)$.

As in Figure 2, for a set of functions constructed on the basis of n pairs of complex poles, the filter consists of n similar sections, each section including two blocks for evaluation of coefficients and one block for the complementary filter.[2]

If the signal input to the spectrum analyzer is $f^*(t)$, it can be expanded as

$$f^*(t) = \sum_{i=0}^{\infty} c_i\phi_i{}^*(t) \tag{15}$$

Thus

$$c_i = \sum_{k=0}^{\infty} f(kT)\phi_i(kT)\ w(kT) \tag{16}$$

where $w(t)$ is the weighting function. Compare Eqns. (14) and (16).

$$f(kT) = m(kT) \tag{17}$$

$$\phi_i(kT)\, w(kT) = g_k(nT-kT) \tag{18}$$

then

$$c_i = h_k(nT) \tag{19}$$

That is, the output of block $\phi_i(z)$ is the coefficient c_i. Such a spectrum analyzer can easily be set up on a digital computer.[2]

Figure 3 represents the block diagram of the basic adaptive discrete system discussed in this paper. It is assumed that the input to the system is a deterministic function of time and that the desired output results from a linear operation on the signal input. It is further assumed that the characteristics of the plant, including the hold if any, vary slowly enough, so that the plant dynamics need be measured only at certain intervals of time. For convenience, this can be chosen to be an integer multiple of the sampling time in the system. Over any such interval called the control interval, the plant dynamics are assumed to be fixed. At the beginning of each interval, the plant characteristics are measured and the controller is adjusted so that some measure of the system error is minimized over the control interval.

The plant characteristics can be determined as above, by feeding the output of the system into a spectrum analyzer. The output from the analyzer will then be the coefficients of the expansion of the plant characteristics in terms of the set of weighted orthonormal functions discussed earlier. The error-function generator operates on the system error to give rise to the measure of error. This information along with the outputs from the spectrum analyzer is fed into a computer which then determines the parameters of the controller.

Stored Energy Response

Since the output of the plant consists of a component caused by present input and one caused by previous inputs, in order to avoid the necessity for a large capacity memory, some way of eliminating the stored energy response has to be adopted.

In any control interval the output, $c(nT)$, of the plant is given in terms of the input as

$$c(nT) = \sum_{k=-\infty}^{n} m(kT)\, g(nT-kT) \tag{20}$$

Divide $c(nT)$ into two parts:

$$c(nT) = \sum_{k=-\infty}^{-1} m(kT)\, g(nT-kT) + \sum_{k=0}^{n} m(kT)\, g(nT-kT) = c_s(nT) + c_m(nT) \tag{21}$$

where time origin is the start of the control interval of interest.

In order to separate the present response c_m from the total response, the output at the end of each control interval can be measured and stored. If each control interval is of length NT, then in the pth control interval, the output $c_m(nT)$, $(p-1)N \leq n \leq pN$, is given as

$$c_m(nT) = c(nT) - c\left[(p-1)NT\right] \tag{22}$$

Alternatively, the stored energy response can be approximated by a set of properly chosen fixed poles, $e^{-a_j T}$, so that the approximate stored energy response $c_{sa}(nT)$ can be written as

$$c_s(nT) = c_{sa}(nT) = \sum_{j=1}^{q} c_j e^{-a_j nT} \tag{23}$$

The c_j and the a_j can be chosen in any appropriate manner. One choice would be to make the actual output and the backward differences of the output up to order q, equal to the approximate response and the backward differences of the approximate response equal at the instant pN-N+1. By a suitable choice of the a_j's, the stored response can be approximated to a high enough degree of accuracy.

The problem now remains to choose the input to the plant, $m(t)$, such that the plant dynamics can be obtained and at the same time the system is made to behave so that the criterion of performance is satisfied. In any control interval, if the input to the plant is $m(nT)$, then the present output (i.e., the component of the output resulting from that part of the input applied during the interval of interest) is given as

$$c_m(nT) = \sum_{k=0}^{n} m(kT)\, g(nT-kT) = \sum_{k=0}^{n} g(kT)\, m(nT-kT),\ 0 \leq n \leq N \tag{24}$$

where $g(t)$ is the impulse response of the plant.

The impulse response may be approximated by an expansion in a set of orthonormal functions $\phi_i(kT)$:

$$g(kT) = g_1\phi_1(kT) + g_2\phi_2(kT) + \dots + g_r\phi_r(kT) \tag{25}$$

The function $c_m(nT)$ can also be approximated by an expansion over the same set of orthonormal functions, so that

$$c_m(nT) = C_1\phi_1(nT) + C_2\phi_2(nT) + \dots + C_r\phi_r(nT) \tag{26}$$

Substitute (25) into (24) and compare with (26). This results in r simultaneous equations involving the coefficients C_i, g_i and the input $m(nT)$. These can be simultaneously solved to give the coefficients g_i in terms of m and C_i. The values of m can easily be obtained from a knowledge of the inputs applied during the previous control interval. The coefficients C_i are determined by passing the signal $c_m(t)$ through the spectrum analyzer. The signal $c_m(t)$ is of course obtained by taking the difference between the total output $c(t)$ and the stored response $c_s(t)$. Since the impulse response of the plant is completely determined by the coefficients g_i, the solution of these r simultaneous equations gives a characterization of the plant dynamics.

In order to complete the problem, it is necessary to choose $m(t)$ such that the actual system output is a good approximation in some sense to the desired output. The system error $e(nT)$ is given as

$$e(nT) = c_d(nT) - c(nT) = \left[c_d(nT) - c_s(nT)\right] - c_m(nT) \tag{27}$$

The output $c_m(nT)$ which satisfies some error criterion, for example zero mean square error over the control interval etc., can be determined from equation (27). Suppose the coefficients of the expansion of this function over the set $\phi_i(kT)$ are C_i. The output at the end of the pth control interval is given as

$$c_m(pNT) = \sum_{k=0}^{pN} g(kT)\, m(pNT-kT) = \sum_{k=0}^{pN} \sum_{i=1}^{r} g_i\phi_i(kT)\, m(pNT-kT) \tag{28}$$

Since this is also equal to $\sum_{i=1}^{r} C_i\phi_i(pNT)$, the values of the input signal $m(nT)$ in the pth control interval can be determined. These values can then be programmed into an input signal generator which will provide the input to the plant. A possible arrangement for setting up the scheme discussed here is shown in block diagram form in Figure 4.

The instrumentation needed for setting up such a system can be considerably simplified if the form of the input forcing function is predetermined. For example, the input could be chosen to have a constant value over each control interval, the magnitude varying from interval to interval.[3] In such a case, over the pth interval, m(t) will have a constant value, say M_p. Then, simultaneous solution of equations (24), (25) and (26) yields simple results for the coefficients g_i as

$$g_i = \frac{C_i}{M_p} \qquad (29)$$

Thus, instead of having to solve r equations simultaneously, the coefficients can be obtained by dividing the output from the spectrum analyzer by the magnitude of the plant input. However, since in this case, there is only one parameter, namely M_k, available for adjusting in each control interval, it is in general not possible to force the system output to be equal to the desired output at all instants in the interval.

References

1. Huggins, William H., "Signal Theory," IRE Transactions on Circuit Theory, Vol. CT-3, No. 4, December 1956.

2. Young, T. Y. and Huggins, W. H. "Discrete Orthonormal Exponentials, Proceedings of the National Electronics Conference, Vol. 18, 1962.

3. Mishkin, Eli and Braun, Ludwig Jr. "Adaptive Control Systems," (book), McGraw-Hill Book Co., New York, 1961.

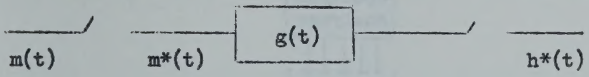


Figure 1

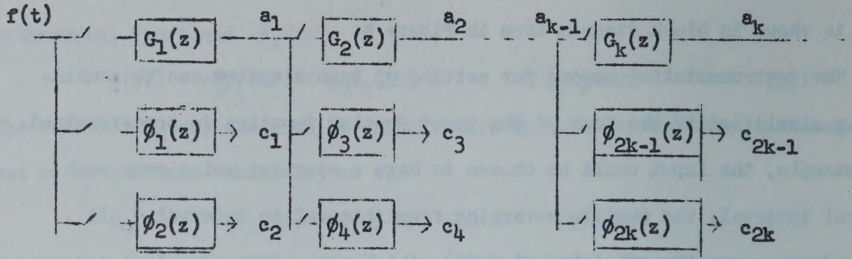

f(t)
$G_1(z)$
a_1
$G_2(z)$
a_2
a_{k-1}
$G_k(z)$
a_k
$\phi_1(z)$
c_1
$\phi_3(z)$
c_3
$\phi_{2k-1}(z)$
c_{2k-1}
$\phi_2(z)$
c_2
$\phi_4(z)$
c_4
$\phi_{2k}(z)$
c_{2k}

Figure 2: The Spectrum Analyzer

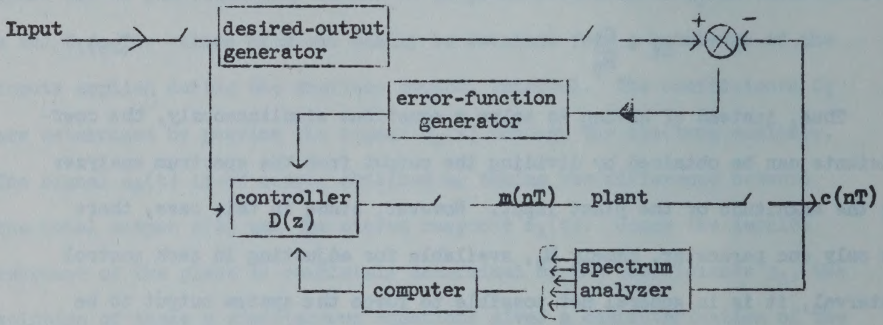

Input
desired-output generator
+
-
error-function generator
controller D(z)
m(nT)
plant
c(nT)
spectrum analyzer
computer

Figure 3: The Basic System

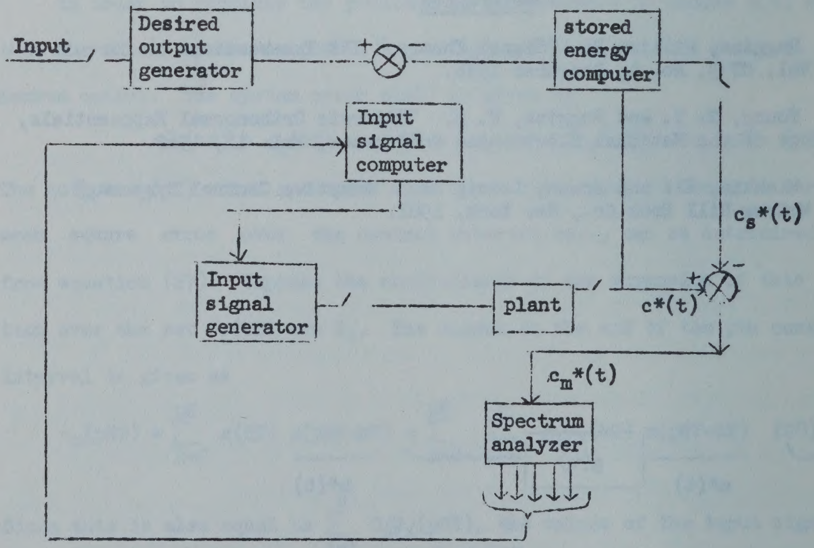

Input
Desired output generator
+
-
stored energy computer
Input signal computer
$c_s*(t)$
Input signal generator
plant
+
-
$c*(t)$
$c_m*(t)$
Spectrum analyzer

Figure 4: The Adaptive Scheme

ON OPTIMUM FINITE DIMMENSIONAL SIGNAL REPRESENTATION

R. E. Totty and J. C. Hancock
School of Electrical Engineering
Purdue University
Lafayette, Indiana

ABSTRACT

The significance of the problem of finding optimum N-tuple signal representations for a given class of signals is pointed out. Solutions and examples are given for two cases: (a) The class of signals has a probability distribution function assigned to it, and the averaged squared error is to be minimized by proper choice of the N approximating signals. (b) The class of signals is finite dimensional (say M) and $N < M$ approximating signals are chosen to minimize the sum of the squared errors from all possible members of the class.

ON THE DRIVING POINT SYNTHESIS PROBLEM

R. E. KALMAN
Center for Control Theory
Research Institute for Advanced Studies (RIAS)
Baltimore 12, Maryland

ABSTRACT

As a byproduct of research on the Problem of Lur'e in stability theory, it was discovered [1] that positive realness of a rational function of the complex variable (or of a matrix of such functions) is equivalent to a set of algebraic relations. These algebraic relations are directly related to the state-variable characterization of the dynamical equations of a network.

The purpose of this paper is to discuss the role of these relations in a direct solution of the driving-point synthesis problem (ideal transformers admitted) and to indicate some directions for future research.

---

[1] R. E. Kalman, Proc. Nat. Acad. Sci. (USA), 49 (1963), 201-205.

THE TIME-DOMAIN SYNTHESIS OF DISTRIBUTIONS

A. H. Zemanian
Department of Engineering Analysis
State University of New York at Stony Brook
Stony Brook, New York

ABSTRACT

A new dimension to the time-domain synthesis problem is proposed, which concerns itself with the approximation of distributions (i.e., generalized functions) by a realizable signal. A solution is presented, several convergence criteria are discussed, and some examples are given. The method appears to be new for function-approximation as well as for distribution-approximation. It also yields a convenient technique for "visualizing" those distributions that cannot be plotted.

INTRODUCTION

An outstanding question in electrical network theory has been the time-domain synthesis problem. Given a prescribed function f(t) of time t, the objective is to develop an approximation technique whereby another function $f_A(t)$, whose Laplace transform $W_A(s)$ is a real rational function with poles in the left-half s-plane and with more poles than zeros, approximates f(t) in some sense. A great deal of attention has been paid to this problem and a number of solutions have been expounded. References [1] to [45] comprise a certainly incomplete list of such pertinent papers. The purpose of this paper is to suggest a new dimension to this problem and to present a solution to it.

Schwartz's distributions [46] and equivalent types of generalized functions [47-51] are now becoming used in various branches of network theory [52-56]. In view of their effect in other physical and mathematical sciences, it seems safe to predict that the use of distribution theory will lead to simplifications and additional results in network theory. For example, in this paper we pose and solve the problem of synthesizing a lumped passive electrical network whose response to a delta function input approximates a prescribed distribution. In view of certain known network synthesis techniques, the issue degenerates into the following approximation problem. "Given a distribution f(t), approximate it by a real finite linear combination of terms of the form

$$1_+(t)\ t^n\ e^{-at}\ \cos\ (bt+c),$$

where $1_+(t)$ is the Heaviside unit step function, n is a nonnegative integer, and a, b, and c are real constants with $a \geq 0$. In any such term, if a = 0, then n = 0. The Laplace transform of the resulting approximation will be a

---

The work reported in this paper was sponsored by the Air Force Cambridge Research Laboratories, Electronics Research Directorate, Communications Theory Branch, Communications Science Laboratory, Bedford, Mass., under contract AF 19(628)2981.

rational function with real coefficients whose poles outnumber its zeros and have nonpositive real parts; furthermore, the imaginary poles are simple. Such a function is realizable as the transfer function of a lumped passive electrical network.

The method that is proposed here is quite simple and yields a convergent approximation to any distribution $f(t)$ whose support is bounded and contained in the open semi-infinite interval, $0 < t < \infty$. Under certain circumstances the technique can be extended to distributions having supports that are unbounded on the right and are contained in the semi-closed interval $0 \leq t < T$. Moreover, even if we restrict ourselves to ordinary functions, the method appears to be new - at least to the author's knowledge.

A byproduct of this technique is that it yields a convenient means of "visualizing" distributions. For example, one crudely thinks of the delta function as a pulse of unit area having an extremely small base and an extremely large height. Similarly, one such positive pulse followed immediately by such a negative pulse is one of the possible approximations for the first derivative of the delta function multiplied by some constant.

In general, however, it has not been clear how to visualize arbitrary distributions. The approximations generated by the method presented here can be used as picturizations of various types of distributions. Several examples of this are given below.

A DELTA-FORMING SEQUENCE

In this discussion we shall freely employ various concepts and results developed in [46]. A brief sketch of the most important ideas is given in the appendix of the technical report 400-55 of [54]. Becuase of space limitations, most of our assertions will not be proven. Such proofs will appear elsewhere.

The approximation procedure, which we shall discuss subsequently, is based upon a sequence $\{g_\nu\}_{\nu=1}^{\infty}$ of functions that converges to the Dirac delta functional $\delta(t)$ in the topology of the space D' of Schwartz's distributions. (We shall refer to such convergence as convergence in D'.) Each term g_ν of this sequence is given by

$$g_\nu(t) = \frac{1}{\nu} \sum_{k=-\nu^p}^{\nu^p} \exp \frac{i2\pi kt}{\nu} \quad (p > 1;\ \nu=1,2,3,\ldots). \qquad (1)$$

(When p is not an integer, it shall be understood that ν^p is replaced by the largest integer that is no larger than the principal-branch value of ν^p. In our examples we shall set $p = 2$.)

Graphs of $g_\nu(t)$ for several choices of ν are shown in Figure 1. It can be seen that each g_ν consists of a row of pulses. As $\nu \to \infty$, each pulse approaches a delta functional; moreover, the pulse at the origin remains there whereas all the others move steadily out toward $t = +\infty$ or $t = -\infty$.

Let $g_{\nu\rho}(t) = e^{-\rho t} g_\nu(t)$, where ρ is a constant. (Hence, $g_{\nu 0} = g_\nu$.) Since

$\{g_\nu\}_{\nu=1}^{\infty}$ converges in D' to $\delta(t)$, $\{g_{\nu\rho}\}_{\nu=1}^{\infty}$ also converges in D' to $\delta(t)$. For, if θ is in D,

$$<g_{\nu\rho},\theta> = < g_\nu,\ e^{-\rho t}\ \theta(t) > \longrightarrow \theta(0). \qquad (3)$$

We shall also make use of the sequence $\{g_{\nu\rho}\}_{\nu=1}^{\infty}$.

The graphs for $g_{\nu\rho}$ appear the same as those for g_ν except that all magnitudes are modulated by $e^{-\rho t}$.

THE APPROXIMATION METHOD

Let us assume for the moment that the distribution f(t), which we shall approximate by a realizable signal, has a bounded support contained in the interval, $\epsilon \leq t < \infty$, where $\epsilon > 0$. An approximation f_ν to f can be obtained by convolving f with some g_ν in the distributional sense. This convolution will certainly have a sense because f is of bounded support.

$$f_\nu = f * g_\nu$$

$$= \frac{1}{\nu} \sum_{k=-\nu^p}^{\nu^p} \left\langle f(\tau),\ \exp \frac{i2\pi k(t-\tau)}{\nu} \right\rangle$$

$$= \frac{1}{\nu} \sum_{k=-\nu^p}^{\nu^p} F(i\frac{2\pi k}{\nu})\ \exp \frac{i2\pi kt}{\nu} \qquad (4)$$

Here, F denotes the distributional Laplace transform of f.

$$F(s) = \left\langle f(\tau),\ e^{-s\tau} \right\rangle$$

Because of the continuity of the convolution process, $\{f_\nu\}_{\nu=1}^{\infty}$ converges in D' to f. Moreover, since f(t) equals the zero distribution over $-\infty > t < \epsilon > 0$, we can truncate the approximation (4) at t = 0 without upsetting the convergence. In short, the sequence $\{f_{\nu+}\}_{\nu=1}^{\infty}$, whose terms are given by $f_{\nu+} = f_\nu(t) \times 1_+(t)$ where 1_+ denotes the Heaviside unit step function, converges in D' to f. Moreover, each approximation $f_{\nu+}(t)$ is realizable by a passive lossless network. The approximation $f_{5+}(t)$ for $f(t) = \delta^{(1)}(t-1)$, where p = 2, is shown in Figure 2. (Henceforth, $h^{(n)}$ shall denote the nth distributional derivative of the distribution h.)

On the other hand, one might wish to approximate by a signal that is realizable by a passive dissipative network. This can be accomplished by convolving f by some $g_{\nu\rho}$, where $\rho > 0$.

$$f_{\nu\rho}(t) = f * g_{\nu\rho}$$

$$= \frac{1}{\nu} \sum_{k=-\nu^p}^{\nu^p} \left\langle f(\tau),\ \exp\ [(-\rho + \frac{i2\pi k}{\nu})(t-\tau)] \right\rangle$$

$$= \frac{1}{\nu} \sum_{k=-\nu^p}^{\nu^p} F(-\rho + \frac{i2\pi k}{\nu})\ \exp\ (-\rho + \frac{i2\pi k}{\nu})\ t \qquad (5)$$

Since f has a bounded support, F(s) is analytic over the entire s-plane. Once the Laplace transform F has been determined (in many cases it can be found in a table) the coefficients $F(-\rho + i2\pi k/\nu)$ and, thereby, the approximation (5) can be written down without any additional computation.

Once again, a realizable approximation can be generated by truncating (5) at t = 0. That is, set

$$f_{\nu\rho +}(t) = f_{\nu\rho}(t)\, 1_{+}(t). \qquad (6)$$

The corresponding Laplace transform is

$$F_{\nu\rho +}(s) = \frac{1}{\nu} \sum_{k=-\nu^p}^{\nu^p} \frac{F(-\rho + i\frac{2\pi k}{\nu})}{s + \rho - i\frac{2\pi k}{\nu}} \quad (\mathrm{Re}\ s > -\rho) \qquad (7)$$

From standard synthesis techniques, we know that there exists some 2-port network composed of only a finite number of resistors, inductors, and capacitors whose response to a delta functional input is (6). Again by the continuity of convolution and by the fact that f has a bounded support in $0 < \epsilon \leq t < \infty$ it follows that $f_{\nu\rho +}$ converges in D' to f.

Let us now remove some of the restrictions on the support of f. First of all, assume that f(t) still has a bounded support but that it extends up to and includes the origin, t = 0. If f is an integrable function in a neighborhood of the origin, then the above procedure will still yield a realizable approximating sequence that converges in D' to f. However, if f is a singular distribution in every neighborhood of the origin (this will be the case, for example, when f has a delta functional or singular point of a pseudofunction at the origin), then an error may be generated when the approximation $f_{\nu\rho}$ is truncated by multiplying it by $1_{+}(t)$. For instance, if $f(t) = \delta(t)$, then the approximation $f_{\nu +}$ equals $g_{\nu}(t)\, 1_{+}(t)$ and as $\nu \longrightarrow \infty$ this converges in D' to $\delta(t)/2$ rather than to $\delta(t)$.

We overcome this difficulty, if we first translate f(t) to the right somewhat before applying our approximation technique. More specifically, instead of approximating f(t) we approximate f(t - x) where the positive number x can be chosen as small as we wish. Since f(t - x) converges in D' to f(t) as $x \longrightarrow \infty$, we can approximate f by choosing ν sufficiently large and x sufficiently small.

We can relax our restriction on the support of f in still another way. Assume that the support of f(t) is bounded on the left at $t = \epsilon > 0$ but unbounded on the right (i.e., it extends infinitely toward $t = \infty$). Then, $f_{\nu\rho} = f * g_{\nu\rho}$ will not in general exist since $g_{\nu\rho}$ also has an unbounded support. Nevertheless, let us formally compute $f_{\nu\rho}$. We get

$$f_{\nu\rho}(t) = \frac{1}{\nu} \sum_{k=-\nu^p}^{\nu^p} \left\langle f(\tau), \exp\,(\rho - i\frac{2\pi k}{\nu})\tau \right\rangle \exp\,(-\rho + i\frac{2\pi k}{\nu})\, t. \qquad (8)$$

If f is such that its distributional Laplace transform F(s) has a region of convergence, Re s > σ_1 where $\sigma_1 < 0$ then for $0 < \rho < -\sigma_1$ the inner products in the right-hand side of (8) will all have a sense. Indeed, the approximation (8) is the same as (5) except that now ρ cannot be chosen larger than $-\sigma_1$. The truncated signal $f_{\nu\rho}(t)\, 1_+(t)$ is once again the realizable approximation to f.

THE USE OF CESARO MEANS

As indicated in Figures 1 and 2 the approximating functions are highly oscillatory. In fact, it turns out that over intervals, where the distribution f is a continuous function, the approximations $f_{\nu\rho}$ may not even converge uniformly. These features can be eliminated if use is made of the Cesaro means.

For example, the first Cesaro mean of the finite sum g_ν, which we shall denote by $g_\nu^{[1]}$, is

$$g_\nu^{[1]}(t) = \frac{2}{\nu}\left[\frac{1}{2} + \sum_{k=1}^{\nu^p} \frac{\nu^p + 1 - k}{\nu^p + 1} \cos\frac{2\pi k t}{\nu}\right]$$

$$= \frac{1}{\nu(\nu^p + 1)}\left[\frac{\sin\frac{(\nu^p + 1)\pi t}{\nu}}{\sin\frac{\pi t}{\nu}}\right]^2$$

More generally, the α - order Cesaro mean ($\alpha = 1, 2, 3, \dots$) of g_ν shall be denoted by $g_\nu^{[\alpha]}$ and is given by

$$g_\nu^{[\alpha]}(t) = \frac{1}{\nu A_n^\alpha} \sum_{k=0}^{n} A_{n-k}^{\alpha-1} \frac{\sin\left[(k + \frac{1}{2})\frac{2\pi t}{\nu}\right]}{\sin\frac{\pi t}{\nu}},$$

where $n = \nu^p$ and

$$A_n^\alpha = \frac{(n+1)\ (n+2)\ \dots\ (n+\alpha)}{\alpha!}$$

Some computation shows that

$$g_\nu^{[\alpha]}(t) = \frac{2}{\nu}\left\{\frac{1}{2} + \sum_{k=1}^{n} \frac{(n+1-k)\ (n+2-k)\ \dots\ (n+\alpha-k)}{(n+1)\ (n+2)\ \dots\ (n+\alpha)} \cos\frac{2\pi k t}{\nu}\right\} \qquad (9)$$

These functions are plotted in Figures 3 and 4 for $\alpha = 1, 2$ and for $\nu = 3, 5, 7$. As before, $p = 2$. Note that the approximating functions becomes smoother while the convergence becomes slower as we proceed to the higher order Cesaro means.

The functions,

$$g_{\nu\rho}^{[\alpha]}(t) = e^{-\rho t} g_\nu^{[\alpha]}(t) \qquad (\rho \geq 0), \qquad (10)$$

will also be used to generate approximations that are realizable by dissipative networks. Figure 5 shows some examples of it for $\alpha = 1$, $\nu = 3, 5, 7$, $p = 2$, and $\rho = 1/2$.

It is a fact that, as $\nu \longrightarrow \infty$, $g_{\nu}^{[\alpha]}(t)$ converges in D' to $\delta(t)$. It immediately follows that $g_{\nu\rho}^{[\alpha]}(t)$ also converges in D' to $\delta(t)$.

Now, if f is once again a distribution whose support is bounded and contained within the psoitive t-axis, then an approximation to f is found, as before, to be

$$f_{\nu\rho}^{[\alpha]}(t) = g_{\nu\rho}^{[\alpha]} * f \tag{11}$$

$$= \frac{1}{\nu} \sum_{k=-n}^{n} \frac{(n+1-|k|)\ \dots\ (n+\alpha-|k|)}{(n+1)\ \dots\ (n+\alpha)} F(-\rho+i\,\frac{2\pi k}{\nu})\ \exp\ (-\rho+i\frac{2\pi k}{\nu})t \tag{12}$$

$(n=\nu^{p}; p>1;\ \alpha = 1, 2, 3, \dots;\ \nu = 1, 2, 3, \dots)$.

For $\rho \geq 0$, a realizable approximation is

$$f_{\nu\rho+}^{[\alpha]}(t) = f_{\nu\rho}^{[\alpha]}(t)\ 1_{+}(t) \tag{13}$$

By the continuity of convolution and the fact that the support of f is bounded and in $\epsilon < t < \infty$ $(\epsilon > 0)$, it follows that (13) converges in D' to f.

As was mentioned above, when one proceeds toward larger α, the approximations become smoother but slower. Another illustration of this is provided by Figures 2, 6 and 7.

FURTHER EXAMPLES: The Approximation of Pseudofunctions

So far, we have only approximated the delta functional and its derivative. Let us now turn to another class of distributions, the pseudofunctions. These distributions are based on Hadamard's concept of the finite part of a divergent integral.

Let

$$f(t) = \mathrm{Fp}\frac{1_{+}(t-1)\ 1_{+}(2-t)}{t-1}$$

Applying our approximating procedure and the first Cesaro mean, we obtain for $\rho = 0$, $\alpha = 1$, and $p = 2$

$$f_{\nu\rho+}^{[1]}(t) = 1_{+}(t)\frac{2}{\nu^{3}+\nu} \sum_{k=1}^{\nu^{2}} (\nu^{2}+1-k) \left[A(k)\ \cos\ \frac{2\pi k(t-1)}{\nu} + B(k)\ \sin\ \frac{2\pi k(t-1)}{\nu}\right].$$

Here,

$$A(k) = \int_{0}^{\frac{2\pi k}{\nu}} \frac{(\cos\ x)-1}{x}\, dx;\quad B(k) = \int_{0}^{\frac{2\pi k}{\nu}} \frac{\sin\ x}{x}\, dx.$$

For $\nu = 11$, this approximation is plotted in Figure 8.

As a more substantial example, consider the pseudofunction,

$$f(t) = \mathrm{Fp}\ \frac{1_{+}(t-1)\ 1_{+}(2-t)}{(t-1)^{3}}.$$

For $\rho = 0$, we obtain in this case the following Cesaro approximating means.

$$F^{[\alpha]}_{\nu\rho+}(t) = 1_+(t) \left\{ -\frac{1}{2\nu} + \frac{2}{\nu} \sum_{K=1}^{\nu^p} C_{k\nu\alpha} \left[(A_{k\nu} - \frac{1}{2}) \cos at + (B_{k\nu} - a) \sin at \right] \right\}$$

Here,

$$a = \frac{2\pi k}{\nu} ,$$

$$A_{k\nu} = \int_o^1 \frac{1}{x^3} \left[(\cos ax) - 1 + \frac{a^2 x^2}{2} \right] dx ,$$

$$B_{k\nu} = \int_o^1 \frac{1}{x^3} \left[(\sin ax) - ax \right] dx ,$$

$$C_{k\nu\alpha} = \frac{(\nu^p + 1 - k) \ldots (\nu^p + \alpha - k)}{(\nu^p + 1) \ldots (\nu^p + \alpha)}$$

This is plotted in Figure 9 for $\alpha = 1, 2, 3$ and $\nu = 5$. Note again the smoothing effect as we proceed to the higher order Cesaro means. These graphs provide convenient visualizations of the distribution in the neighborhood of its singular point.

A USEFUL CONVERGENCE PROPERTY

This approximation technique possesses a certain convergence property that makes it useful not only for distributions but also for ordinary functions. In stating this property we shall need the concept of the order of a distribution. A distribution f is said to be of finite order if there exists a non-negative integer q such that f(t) is the qth distributional derivative for all t of a continuous function h(t); (i.e., $f(t) = h^{(q)}(t)$). In this case, f is said to be of order r (r = 0, 1, 2, ...) where r is the smallest of the non-negative integers q for which $f(t) = h^{(q)}(t)$. The said convergence property is the following.

If f(t) is of bounded support and of order r (r = 0, 1, 2, ...) and if it is a continuous function over $a < t < b$, then $f^{[\alpha]}_{\nu\rho}(t)$ converges uniformly to f(t) over $c \leq t \leq d$, where $a < c < d < b$, so long as $\alpha > r$.

A proof of this assertion will appear elsewhere.

ACKNOWLEDGEMENT

The author is indebted to Mr. Timothy Loughlin who performed all the numerical computations and prepared all the graphs that are presented in this paper.

REFERENCES

1. Guillemin, E.A.: Synthesis of Passive Networks. John Wiley and Sons, New York, Ch. 15, 1957.

2. Lee, Y.W.: Synthesis of Electric Networks by Means of the Fourier Transforms of Laguerre Functions. J. of Math. and Phys., 11:83 1931-32.

3. Aigrain, P.R. and Williams, E.M.: Design of Optimum Transient Response Amplifiers. Proc. IRE, Vol. 37 (August) 1949.

4. Aigrain, P.R. and Williams, E.M.: Synthesis of n-reactance Networks for Desired Transient Response. J. of Applied Physics, 20:597-600 (June) 1949.

5. Tuttle, D.F. Jr.: Network Synthesis for Prescribed Transient Response. D. Sc. Thesis, MIT, EE Dept., 1948.

6. Kautz, W.H.: Transient Synthesis in the Time Domain. IRE Trans. on Circuit Theory, Vol. CT-1, pp. 29-39 (Sept.) 1954.

7. Pucel, R.A.: Network Synthesis for a Prescribed Impulse Response Using a Real Port Approximation. J. of Appl. Phys., 28:124-29 (Jan.) 1957.

8. Strieby, M.: A Fourier Method for Time Domain Synthesis. Proc. Symp. Modern Net. Syn., P.I.B. pp. 197-209 (April) 1955.

9. Head, J.W.: Approximations to Transients by Means of Laguerre Series. Proc. Cambridge Phil. Soc., 52:640-51 (Oct.) 1956.

10. Armstrong, H.L.: On the Representation of Transients by Series of Orthogonal Functions. IRE Trans. on Circuit Theory, Vol. CT-6, pp. 351-354 (Dec.) 1959.

11. Yengst, W.C.: Approximation to a Specified Time Response. IRE Trans. on Circuit Theory, Vol. CT-9, pp. 152-62 (June) 1962.

12. Liu, B.: A Time-domain Approximation Method and its Application to Lumped Delay Lines. IRE Trans. on Circuit Theory, Vol. CT-9, pp. 256-261 (Sept.) 1962.

13. Liu, B. and Yengst, W.C.: Approximation to a Specified Time Response. IRE Trans. on Circuit Theory, Vol. CT-10, p. 115 (March) 1963.

14. Ba Hli, F.: A General Method for Time-domain Network Synthesis. IRE Trans. on Circuit Theory, Vol. CT-1, pp. 21-28 (Sept.) 1954.

15. Brule, J.D.: Improving the Approximation to a Prescribed Time Response. IRE Trans. on Circuit Theory, Vol. CT-6, pp. 355-61 (Dec.) 1959.

16. Brule, J.D.: Comments on the Approximation Procedure. IRE Trans. on Circuit Theory, Vol. CT-8, pp. 81-82 (Mar.) 1961.

17. Carr, J.W.: An Analytic Investigation of Transient Synthesis by Exponentials. S.M. Thesis, MIT, EE Dept., 1949.

18. Cerillo, M.V. and Guillemin, E.A.: Rational Fraction Expansions for Network Functions. Proc. Symposium on Modern Network Synthesis. Polytechnic Inst. of Brooklyn, pp. 84-127 (Apr.) 1952.

19. Cruickshank, A.J.O.: Transient Responses with Limited Overshoot-Approximation by Principal Mode. Wireless Engineer, 32:154-63 (June) 1955.

20. D'Atri, M.L.: A Method of Synthesis of Linear Passive Networks for Transient Conditions. Note Recensioni Notiz, 8:155-168 (March-April) 1959.

21. Diamond, J. and Gerst, I.: The Elimination of Intersymbol Interference by Input Signal Shaping. Proc. IRE, 49:1195-1203 (July) 1961.

22. Dolezal, V. and Janac, K.: Synthesis of a Finite Four-terminal Network with a Prescribed Transient Voltage Response. Acta Technica, pp. 479-89, 1957.

23. Elmore, W.C.: The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifiers. J. of Appl. Phys., 19:55-63 (Jan.) 1948.

24. Fawwaz, R.A.: A Design Procedure for Pulse Forming Networks. S.M. Thesis, MIT, E.E. Dept., 1947.

25. Fuller, A.T.: The Adjustment of Control Systems for Quick Transient Response. Proc. I.E.E., Part B., 102:596-601 (Sept.) 1955.

26. Ganguli, R.C.: Design of Pulse Amplifiers. Indian J. of Phys., 33:263-75 (June) 1959.

27. Graham, D. and Lathrop, R.C.: The Synthesis of Optimum Transient Responses: Criteria and Standard Forms. Trans. AIEE, Part II, Applications and Industry, 72:273-88 (Nov.) 1953.

28. Hoffman, R. and Walcher, W.: The Influence of the Coefficients of the Transfer Function of a Transmission System on the Output Characteristics as a Function of Time. Arch. Elek. Ubertragung, 11:321-24 (Aug.) 1957.

29. Huggins, W.H.: Network Approximation in the Time Domain. Air Force Cambridge Research Laboratories, Report E5048A, 1950.

30. Kallman, H.E., Spencer, R.E., and Singer, C.P.: Transient Phenomena. Proc. IRE, 33:169-96 (Mar.) 1945.

31. Kochanov, N.S.: On a Method of Synthesizing Linear Electrical Networks from Specified Time Characteristics. Izvestia V.U.Z., Radiotekhnika, pp. 321-32 (May-June) 1959.

32. Krize, S.N.: The Design of Complex Pulse Systems for Given Transient Characteristics. Radio Engineering, 13:47-51 (Sept.) 1958.

33. Kulikowski, R.: Optimum Equivalent Networks and Models of Dynamic Linear Systems. Archiqum Elektrotechniki, 4:17-34, 1955.

34. Kulikowski, R.: Determination of the Optimum Linear Network Parameters Based on Time Characteristics. Archiwum Elektrotechniki, 4:323-46, 1955.

35. Marsten, R.B.: Synthesis of Video Amplifiers for Prescribed Transient Responde. S.M. Thesis, Dept. of Elect. Engr., MIT, 1946.

36. Mathers, G.W.C.: The Synthesis of Lumped Element Circuits for Optimum Transient Response. Electronics Research Lab., Stanford University, Tech. Rep. No. 38 (Nov.) 1951.

37. Matkhanov, P.N.: The Synthesis of Linear Pulse Shaping Networks by Expanding Integral Functions into Infinite Products. Izvestia V.U.Z., Radiotekhnika, pp. 58-64 (Jan-Feb) 1959.

38. Mulligan, J.H., Jr.: The Effect of Pole and Zero Locations on the Transient Response of Linear Dynamic Systems. Proc. IRE, Vol. 37 (May) 1959.

39. Nadler, M.: The Synthesis of Electric Networks According to Prescribed Transient Calculations. Proc. IRE, 37:627-30 (June) 1949.

40. Reza, F.M.: An Introduction to a Qualitative Synthesis of Linear Passive Networks in the Time Domain. Proc. National Electronics Conference, Vol. 10 (Feb.) 1955.

41. Ruston, H.: Synthesis of RLC Networks by Discrete Tschebyscheff Approximations in the Time-domain. Doctoral Thesis, University of Michigan, 1960.

42. Smith, G.W., Jr.: Dual Requirements for Optimizing Orthonormal Approximations for Impulse Responses. IRE Trans. on Circuit Theory, Vol. CT-8, pp. 486-87 (Dec.) 1961.

43. Thomson, W.E. and Elmore, H.J.: Optimum Coupling Networks. IRE Trans. on Circuit Theory, Vol. CT-4, pp. 55-56 (June) 1957.

44. White, W.F.: Application of Transitional Transients to Network Design. S.M. Thesis, MIT, E.E. Dept., 1950.

45. Zemanian, A.H.: Further Effects of the Pole and Zero Locations on the Step Response of Fixed Linear Systems. Trans. of the A.I.E.E., Communications and Electronics (March) 1955, pp. 52-55.

46. Schwartz, L.: Theorie des Distributions, Vol. I and II. Actualites Scientifiques et Industrielles, Herman, Paris, 1957 and 59.

47. Lighthill, M.J.: Fourier Analysis and Generalized Functions. Cambridge University Press, London, 1958.

48. Gelfand, I.M. and Shilov, G.E.: Generalized Functions, Vols. 1 through 3. Moscow, 1958, 1959.

49. Mikusinski, J.: Operational Calculus. Pergamon Press, New York, 1959.

50. Konig, H.: Neue Begrundung der Theorie der "Distributionen" von L. Schwartz. Math. Nachr., 9:129-48, 1953.

51. Sauer, R.: Aufangswertprobleme bei Partiellen Differentialgleishungen. Second Ed., Springer Verlag O.H.G., Berlin, 1958.

52. Dolezal, V.: Uber die "Wiederbelebung" der Fourierschen Transformation durch die Distributionentheorie. Acta Technica, pp. 335-55, 1961.

53. Dolezal, V.: Uber die Anwendung von Operatoren in der Theorie den Linearen Dynamischen Systeme. Applikace Matematiky, 6:36-67, 1961.

54. Zemanian, A.H.: An n-Port Realizability Theory based on the Theory of Distributions. Tech Rep. 400-55, Dept. of Elect. Engrg., New York University (March) 1962. Also, I.E.E.E. Trans. on Circuit Theory, Vol. CT-10 (June, 1963).

55. Zemanian, A.H.: Distributional Convolution and an Application to Network Realizability Theory. Proceedings of the Sixth Midwest Symposium on Circuit Theory, University of Wisconsin, Madison, Wisconsin (May 6-7) 1963.

56. Newcomb, R.W.: The Foundations of Network Theory. Tech. Rep. No. 2250-4, Stanford Electronics Lab., Stanford University, Stanford, California (March) 1963.

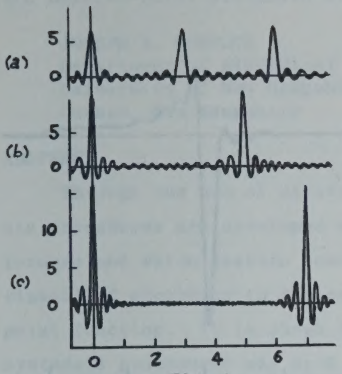


FIG. 1

(a) $\nu = 3$. (b) $\nu = 5$ (c) $\nu = 7$

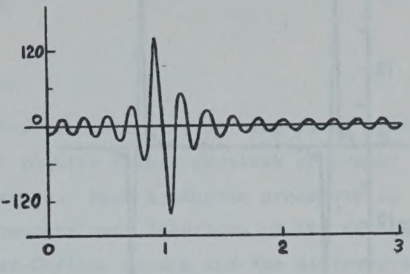


FIG. 2

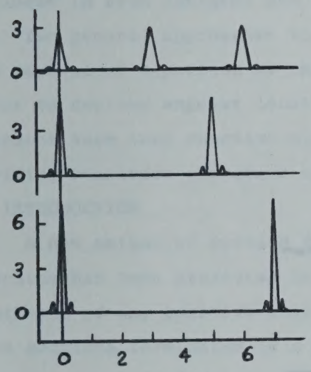


FIG. 3

(a) $\nu = 3$, $\alpha = 1$. (b) $\nu = 5$, $\alpha = 1$. (c) $\nu = 7$, $\alpha = 1$.

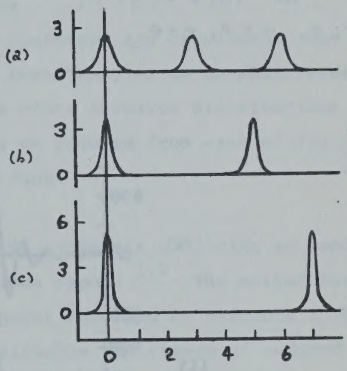


FIG. 4

(a) $\nu = 3$, $\alpha = 2$, (b) $\nu = 5$, $\alpha = 2$. (c) $\nu = 7$, $\alpha = 2$.

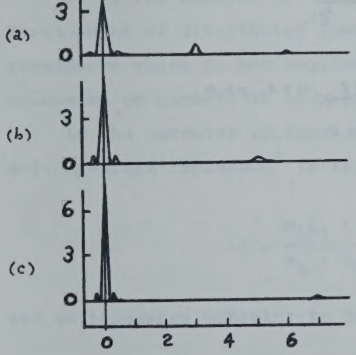


FIG. 5

(a) $\nu = 3$, $\alpha = 1$, $\rho = 1/2$. (b) $\nu = 5$, $\alpha = 1$, $\rho = 1/2$.
(c) $\nu = 7$, $\alpha = 1$, $\rho = 1/2$.

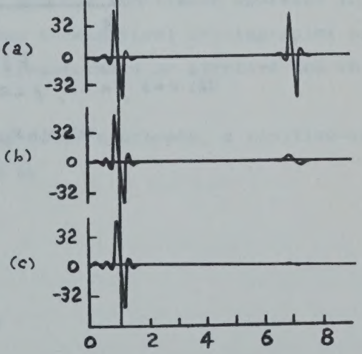


FIG. 6

$f^{[\alpha]}_{\nu\rho+}(t)$ for $f(t) = \delta^{(1)}(t-1)$

(a) $\nu = 5$, $\alpha = 1$, $\rho = 0$. (b) $\nu = 5$, $\alpha = 1$, $\rho = 1/2$.
(c) $\nu = 5$, $\alpha = 1$, $\rho = 1$.

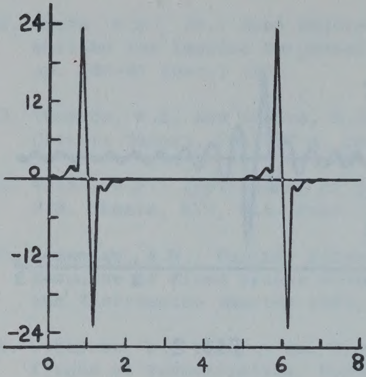


FIG. 7

$f^{[\alpha]}_{\nu\rho+}(t)$ for $f(t) = \delta^{(1)}(t-1)$, $\nu = 5$, $\alpha = 2$, $\rho = 0$.

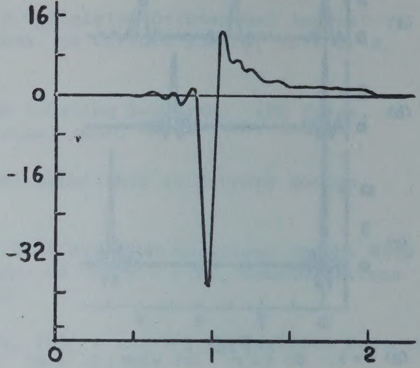


FIG. 8

$\nu = 11$, $\alpha = 1$, $\rho = 0$.

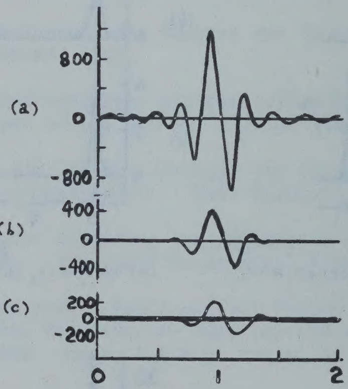


FIG. 9

(a) $\nu = 5$, $\alpha = 1$, $\rho = 0$. (b) $\nu = 5$, $\alpha = 2$, $\rho = 0$

(c) $\nu = 5$, $\alpha = 3$, $\rho = 0$.

RLC DRIVING-POINT SYNTHESIS BY DISTRIBUTED OPERATORS*

JOSEPH B. MURDOCH
Department of Electrical Engineering
University of New Hampshire
Durham, New Hampshire

ABSTRACT

Through the use of distributed immittance operators, driving-point synthesis procedures are developed which do not require either gyrators or transformers and which feature economy of elements. Each synthesis procedure is classified according to the permitted even-part zero locations of the driving-point function. It is shown that the Bott-Duffin, Miyata and Kuh distributed synthesis procedures may each be utilized in the immittance operator method and that a variety of network topologies may be obtained in the synthesis of a given immittance. This latter feature is largely due to the arbitrary constants included in each operator and termination.

Two general approaches to operator synthesis are developed. One features the removal of a portion of the minimum resistance so as to shift even part zeros to desired angular locations. The other involves distributions of the operator such that reactive elements may be removed from each of its parts, yielding remainder operators of reduced rank.

I INTRODUCTION

A new method of cascade driving-point synthesis utilizing an immittance operator has been presented in two previous papers.[1,2] The method permits a synthesis of any positive real driving-point function by cascaded lossless network sections terminated in a single realizable immittance of reduced rank.[3] However, because of the constraint of a single termination, it is usually necessary to include ideal gyrators or perfect transformers in the resulting networks.

It is the purpose of this paper to use the immittance operator in the development of distributed (more than one termination) driving-point synthesis procedures which do not require either transformers or gyrators and which economize on numbers of elements.

In the operator approach to driving-point synthesis, a positive-real driving-point impedance[4] is represented by

$$Z = \frac{m_1\zeta_1 + n_1}{m_2 + n_2\zeta_1} \tag{1}$$

and an impedance operator is defined by

$$V_1 = \frac{m_1 + n_1}{m_2 + n_2} \tag{2}$$

In this representation, m_1 and m_2 are even polynomials in s, n_1 and n_2 are odd polynomials in s, ζ_1 and V_1 are positive real impedances and V_1 is further restricted to represent a lossless network terminated in a one-ohm resistance.

The operator representation can be extended by letting

$$\zeta_1 = \frac{m_3\zeta_2 + n_3}{m_4 + n_4\zeta_2} \tag{3}$$

and

$$V_2 = \frac{m_3 + n_3}{m_4 + n_4} \tag{4}$$

By continuing the process, Z can be written in operational form as

$$Z = V_1V_2 \ldots V_n\zeta_n \tag{5}$$

and a synthesis of any positive-real driving-point impedance can be achieved as shown in Figure 1.[5] In Figure 1, each V box is a lossless network containing one or more arbitrary constants. Furthermore the following relation holds:[6]

$$\text{num EvZ} = (\text{num EvV}_1)\ (\text{num EvV}_2) \ldots (\text{num EvV}_n)\ (\text{num Ev}\zeta_n) \tag{6}$$

where "num Ev" means numerator of the even part. If the arbitrary constants are all chosen to achieve rank reduction in $\zeta_1, \zeta_2 \ldots \zeta_n$ (zero-cancellation synthesis),[7] then each V box realizes one or more sets of even-part zeros of Z. Use of an arbitrary constant for other purposes (such as simplifying the realization of a V box) effectively multiplies Z by a surplus factor. Then the particular V box realizes the even-part zeros of the surplus factor.

The procedure just described is a cascade one in that only a single termination is used. This constraint of a single termination can be removed by noting that the V operator possesses the properties that

$$V\zeta = (V_a + V_b + \ldots)\ \zeta \tag{7}$$

and

$$\text{num EvV} = \text{num EvV}_a + \text{num EvV}_b + \ldots \tag{8}$$

provided that the denominators of V, V_a, V_b, ... are identical. Furthermore, if the distribution of V is made such that V_a, V_b, ... are themselves positive real functions, then each may be separately synthesized with a one-ohm resistive termination. Replacing each resistance by ζ (sometimes $1/\zeta$) and assembling the resulting networks in series yields Z. This is a distributed (as opposed to a cascade) synthesis procedure, as indicated in Figure 2 (number subscripts are used for cascaded operators and letter subscripts for distributed operators).

II SPECIFIC OPERATORS

The immittance operator approach is successful only if a given Z can be separated into two or more parts V_1, V_2, ... V_n, ζ_n. Specific immittance operators are needed. In this paper the distributed synthesis of bilinear, biquadratic, and bicubic operators is considered. To develop these operators, define the following bilinear transformations:

$$Z = \frac{a\zeta_1 + sZ_a}{a + \frac{s}{Z_a}\zeta_1} \quad \text{and} \quad V_1 = \frac{a + sZ_a}{a + \frac{s}{Z_a}} \tag{9}$$

$$\zeta_1 = \frac{b\zeta_2 + s\zeta_{1b}}{b + \frac{s}{\zeta_{1b}}\zeta_2} \quad \text{and} \quad V_2 = \frac{b + s\zeta_{1b}}{b + \frac{s}{\zeta_{1b}}} \tag{10}$$

$$\zeta_2 = \frac{c\zeta_3 + s\zeta_{2c}}{b + \frac{s}{\zeta_{2c}}\zeta_2} \quad \text{and} \quad V_3 = \frac{c + s\zeta_{2c}}{c + \frac{s}{\zeta_{2c}}} \tag{11}$$

Z_a, ζ_{1b} and ζ_{2c} are the values of Z, ζ_1 and ζ_2 at s = a, b and c, respectively. V_1 is a bilinear operator relating Z to ζ_1 by

$$Z = V_1\zeta_1 \tag{12}$$

According to Richards' Theorem,[8] ζ_1 and V_1 are positive real if Z is positive real and if a is a positive real constant.

To develop a biquadratic operator, let V_1 operate on V_2 (in the same way that V_1 operates on ζ_1 in (1)) to yield

$$V_{12} \equiv V_1V_2 = \frac{ab + Ns + \frac{Z_a}{\zeta_{1b}}s^2}{ab + Ds + \frac{\zeta_{1b}}{Z_a}s^2} \tag{13}$$

where

$$N = bZ_a + a\zeta_{1b}, \quad D = \frac{b}{Z_a} + \frac{a}{\zeta_{1b}} \tag{14}$$

In operational form, Z is now given by

$$Z = V_1V_2\zeta_2 \equiv V_{12}\zeta_2 \tag{15}$$

From an extension of Richards' Theorem by Hazony,[9] ζ_2 and V_{12} are positive real if Z is positive real and if a and b are either positive real or complex conjugates with a non-negative real part.

The last operator we wish to develop is a bicubic operator, which is obtained by letting V_{12} operate on V_3. The result is

$$V_{123} \equiv V_{12}V_3 = \frac{abc + N_1 s + N_2 s^2 + \frac{Z_a \zeta_{2c}}{\zeta_{1b}} s^3}{abc + D_1 s + D_2 s^2 + \frac{\zeta_{1b}}{Z_a \zeta_{2c}} s^3} \tag{16}$$

where

$$N_1 = ab\zeta_{2c} + bcZ_a + ca\zeta_{1b}, \quad N_2 = a\frac{\zeta_{1b}}{\zeta_{2c}} + b\frac{Z_a}{\zeta_{2c}} + c\frac{Z_a}{\zeta_{1b}}$$
$$D_1 = \frac{ab}{\zeta_{2c}} + \frac{bc}{Z_a} + \frac{ca}{\zeta_{1b}}, \quad D_2 = a\frac{\zeta_{2c}}{\zeta_{1b}} + b\frac{\zeta_{2c}}{Z_a} + c\frac{\zeta_{1b}}{Z_a} \tag{17}$$

In operational form, Z is now given by

$$Z = V_1V_2V_3\zeta_3 = V_{12}V_3\zeta_3 = V_{123}\zeta_3 \tag{18}$$

If ζ_3 and V_{123} are each to be positive real, one of the constants (a, b, or c) must be positive real while the other two may be either positive real or complex conjugates with a non-negative real part.[10]

III BILINEAR AND BIQUADRATIC OPERATOR SYNTHESIS

The bilinear operator V_1 in (9) has an even part given by

$$\mathrm{Ev}V_1 = \frac{a^2 - s^2}{a^2 - \frac{1}{Z_a^2} s^2} \tag{19}$$

$\mathrm{Ev}V_1$ may be split into the following two parts according to (7) and (8)

$$\mathrm{Ev}V_1 = \frac{a^2}{a^2 - \frac{1}{Z_a^2} s^2} + \frac{-s^2}{a^2 - \frac{1}{Z_a^2} s^2} \tag{20}$$
$$= \mathrm{Ev}V_{1a} + \mathrm{Ev}V_{1b}$$

and each of the parts may be synthesized by either the Miyata or Darlington[11] procedures to yield the network of Figure 3.

If Z has a set of real even part zeros (num EvZ contains a term $\sigma_o^2 - s^2$), choosing $a = \sigma_o$ will permit V_1 to realize the set with two reactive elements and make ζ_1 two less in rank than Z. Alternatively, if Z is minimum resistive, a may be chosen to create a removable resonant circuit in ζ_1 (as in the Bott-

Duffin[12] procedure). This provides a rank reduction of four but requires six reactive elements. Thus the bilinear operator can be used to realize real and imaginary, but not complex, even part zeros.

Now consider the biquadratic operator in (13). Its even part is given by (6) as

$$\mathrm{Ev}V_{12} = \frac{(a^2 - s^2)(b^2 - s^2)}{P^2 - Q^2} \tag{21}$$

where P is the even part of the denominator in (13) and Q is its odd part.

A. Kuh-Split

The immediate task is to determine ways in which V_{12} (or $\mathrm{Ev}V_{12}$) may be distributed such that each of its component parts is a realizable impedance. To this end, (21) is rewritten in the distributed form

$$\mathrm{Ev}V_{12} = \frac{a^2b^2}{P^2 - Q^2} + \frac{-(a^2 + b^2)s^2}{P^2 - Q^2} + \frac{s^4}{P^2 - Q^2} = \mathrm{Ev}V_{12a} + \mathrm{Ev}V_{12b} + \mathrm{Ev}V_{12c} \tag{22}$$

The first and third terms in (22) represent realizable impedances for any permissible a and b. The second term does only if

$$\text{Real } a \geq |\text{Imag. } a| \tag{23}$$

If Z has a quadruplet of even part zeros within 45° of the real axis, a and b may be chosen such that (23) is satisfied. Then V_{12} realizes the quadruplet and ζ_2 is four less in rank than Z.

Each of the terms in (22) contains a single power of s. A saving in elements occurs if the second term is combined with either the first or third terms. This will be termed a Kuh split.[13]

Combining the second and third terms yields

$$\mathrm{Ev}V_{12} = \frac{a^2b^2}{P^2 - Q^2} + \frac{-(a^2 + b^2)s^2 + s^4}{P^2 - Q^2} \equiv \mathrm{Ev}V_{12a} + \mathrm{Ev}V_{12bc} \tag{24}$$

The synthesis of the two V_{12} parts is accomplished by the Miyata procedure. V_{12a} and V_{12bc} are found from their even parts to be

$$V_{12a} = \frac{ab + s\dfrac{ab\zeta_{1b}}{DZ_a}}{ab + sD + s^2\dfrac{\zeta_{1b}}{Z_a}}\ , \quad V_{12bc} = \frac{\dfrac{S}{D}\left(a^2 + b^2 + ab\dfrac{Z_a}{\zeta_{1b}}\right) + s^2\dfrac{Z_a}{\zeta_{1b}}}{ab + sD + s^2\dfrac{\zeta_{1b}}{Z_a}} \tag{25}$$

$V_{12a}\zeta_2$ can be synthesized directly by a Cauer expansion

$$V_{12a}\zeta_2 = \cfrac{1}{s\dfrac{D}{ab} + \cfrac{1}{s\dfrac{\zeta_{1b}}{Z_aD} + \zeta_2}} \tag{26}$$

$V_{12bc}\zeta_2$ must be synthesized in steps. First the zero at the origin is removed giving

$$V_{12bc}\zeta_2 = \cfrac{1}{\cfrac{abD}{s(a^2 + b^2 + ab\frac{Z_a}{\zeta_{1b}})} + \cfrac{1}{V_r\frac{1}{\zeta_2}}} \tag{27}$$

where

$$V_r = \frac{\frac{1}{D}(a^2 + b^2 + ab\frac{Z_a}{\zeta_{1b}}) + s\frac{Z_a}{\zeta_{1b}}}{\frac{(a^2 + b^2)D}{a^2 + b^2 + ab\frac{Z_a}{\zeta_{1b}}} + s\frac{\zeta_{1b}}{Z_a}} \tag{28}$$

and

$$\text{num Ev}V_r = (a^2 + b^2) - s^2 \tag{29}$$

Two alternatives are available at this point. V_r is a bilinear operator and is realizable if (23) is satisfied. The synthesis of V_r yields a network having the form of Figure 3, and the synthesis of Z is then as shown in Figure 4a, where the K's are scaling constants. Three terminations, each four less in rank than Z, and five reactive elements are required. If the even part zeros lie at 45^o, $a^2 + b^2 = 0$ and only four reactive elements and two terminations are needed.

As an alternative to synthesizing V_r, we can recombine this operator with ζ_2. Then the synthesis of Z is given by Figure 4b. Only two terminations, one two less in rank than Z and the other four less in rank than Z, and three reactive elements are required. This latter procedure, in which a termination and a remainder operator are recombined, often leads to a reduction in total number of elements, especially if the terminations are of high rank requiring further cycles in the synthesis.

We noted that, when economy of elements is important, a Kuh split is preferable to the split in (22). In general, a distribution in which num Ev V is separated into two parts having equal (or nearly equal) numbers of terms, with all high-degree terms in one part and low-degree terms in the other, is optimum. Such a distribution is known as a Half-N split.[14]

B. Resonant-Circuit Split

The biquadratic operator synthesis just discussed is unsuccessful in reducing the rank of ζ_2 when the even part zeros in question are more than 45^o from the real axis. To handle such cases, reconsider V_{12} in (13) and assume a and b are such that $Z_{a/\zeta_{1b}} > 1$. Let V_{12} be distributed in the following way.

$$V_{12} = \frac{ab + s^2 \frac{\zeta_{1b}}{Z_a}}{ab + sD + s^2 \frac{\zeta_{1b}}{Z_a}} + \frac{sN + s^2 \left(\frac{Z_a}{\zeta_{1b}} - \frac{\zeta_{1b}}{Z_a}\right)}{ab + sD + s^2 \frac{\zeta_{1b}}{Z_a}} \equiv V_{12a} + V_{12b} \tag{30}$$

V_{12a} is certainly positive real. The coefficients of V_{12b} are positive real. V_{12b} has an even part numerator given by

$$\text{num Ev}V_{12b} = s^4\left(1 - \frac{\zeta_{1b}^2}{Z_a^2}\right) - s^2\left(a^2 + b^2 + 2\,ab\,\frac{\zeta_{1b}}{Z_a}\right) \tag{31}$$

The s^4 coefficient is positive. To determine whether the $-s^2$ coefficient is also positive, let

$$a = x + jy,\ b = x - jy \tag{32}$$

Then the second term in (31) yields the following sufficient condition for positive realness.

$$\frac{x^2\left(1 + \frac{\zeta_{1b}}{Z_a}\right)}{y^2\left(1 - \frac{\zeta_{1b}}{Z_a}\right)} \geq 1 \tag{33}$$

This result may be simplified considerably. Evaluating ζ_1 in (9) at $s = b$, substituting in (33), and letting

$$Z_a = u + jv,\ Z_b = u - jv \tag{34}$$

gives

$$y^2 \leq x^2 \frac{y/x}{v/u} \tag{35}$$

Since the argument of Z_a must be less than or equal to the argument of a, (35) becomes

$$y^2 \leq K^2x^2,\ K \geq 1 \tag{36}$$

which is always satisfied for even-part zeros within 45° of the real axis and is often satisfied when they are not. Thus the Resonant Circuit split is more general than the Kuh split. For example, if $K^2 = 3$, then $y^2 \leq 3X^2$, $y \leq \sqrt{3}X$, and a quadruplet of even part zeros up to 60° from the real axis will still keep V_{12b} positve real.

If $Z_{a/\zeta_{1b}} < 1$, v is negative and a similar development yields the positive real condition of (35) with v replaced by -v.

The synthesis of $V_{12}\zeta_2$ (V_{12} given by (30)) proceeds in the same way as the

Kuh split synthesis. Again a choice is available in the synthesis of V_{12b}. If this operator is completely synthesized, Figure 5a results. Five reactive elements and three terminations, each four less in rank than Z, are required. Alternatively the shunt inductance may be removed from V_{12b} and the remainder operator recombined with ζ_2. Three reactive elements and two terminations, one four less in rank than Z and the other two less are needed. The network appears in Figure 5b.

If (35) is satisfied with the equal sign (vy = ux), the $-s^2$ coefficient in (31) vanishes and $V_{12}\zeta_2$ can be synthesized with 4 reactive elements and 2 terminations. Since $\frac{v}{u}$ must be less than $\frac{y}{x}$, the equal sign can apply only for even part zeros more than 45° from the real axis.

C. Even-Odd Split

The Resonant Circuit split, while more general than a Kuh Split, is nevertheless not always successful in realizing even part zeros more than 45° from the real axis. A completely general procedure, involving the removal of a portion of the odd part of the biquadratic operator, will now be developed.

Let EvV_{12} in (21) be distributed to yield

$$EvV_{12} = \frac{P_o^2}{P^2 - Q^2} + \frac{-Q_o^2}{P^2 - Q^2} = \frac{(ab + s^2)^2}{P^2 - Q^2} + \frac{-(a + b)^2 s^2}{P^2 - Q^2} \equiv EvV_{12a} + EvV_{12b} \qquad (37)$$

where P_o and Q_o are even and odd polynomials in s, respectively.

Both EvV_{12a} and EvV_{12b} represent realizable impedances. Essentially a portion of the odd part of the numerator of V_{12} has been removed and assigned to V_{12b} so that V_{12a} has imaginary even part zeros. This is more economical in elements than removing V_{12}'s minimum resistance to accomplish the same result.[15]

$V_{12b}\zeta_2$ can be synthesized directly by two reactive elements and a termination of $\frac{(a + b)^2}{D^2\zeta_2}$. To synthesize $V_{12a}\zeta_2$ without coupling, replace ζ_2 by $V_3\zeta_3$ from (11).

Then

$$V_{12a}\zeta_2 = V_{12a}V_3\zeta_3 \qquad (38)$$

and

$$EvV_{12a}V_3 = \frac{(s^2 + ab)^2\ (c^2 - s^2)}{(P^2 - Q^2)(c^2 - \frac{s^2}{\zeta_{2c}^2})} \qquad (39)$$

$EvV_{12a}V_3$ can be split into two parts:

$$EvV_{12a}V_3 = \frac{c^2(s^2 + ab)^2}{(P^2 - Q^2)(c^2 - \frac{s^2}{\zeta_{2c}^2})} + \frac{-s^2(s^2 + ab)^2}{(P^2 - Q^2)(c^2 - \frac{s^2}{\zeta_{2c}^2})} \equiv EvV_x + EvV_y \qquad (40)$$

EvV_x and EvV_y represent realizable impedances. Each may be synthesized without coupling, since c may always be chosen to satisfy one of the following two conditions:[16]

$$c\zeta_{2c} = \frac{D}{1 - \frac{\zeta_{1b}}{Z_a}} \qquad \text{or} \qquad c\zeta_{2c} = \frac{ab}{D}\left(\frac{\zeta_{1b}}{Z_a} - 1\right) \tag{41}$$

The resulting network when the first condition is satisfied appears in Figure 6 and contains eight elements and three terminations, or three more elements per cycle than the network of Figure 5. If Z is minimum resistive, or is made so by first removing its minimum resistance, the upper network in Figure 6 vanishes $(a + b = 0)$. The remaining network is identical to that which would be obtained by using the Bott-Duffin procedure to synthesize Z.

IV BIQUADRATIC OPERATOR SUMMARY AND EXAMPLE

When the quadruplet of even part zeros being realized is less than 45° from the real axis, Procedure A should be used. A synthesis with 5 elements and 3 terminations results.

If the zeros are at 45°, Procedure A should be used and yields a synthesis with 4 elements and 2 terminations $(a^2 + b^2 = 0)$.

For zeros between 45° and 90° from the real axis, Procedure B should be tried. If it succeeds, a synthesis with 5 elements and 3 terminations is achieved. If it fails, Procedure C is used, resulting in a synthesis with 8 elements and 3 terminations (7 elements and 2 terminations if Z's minimum resistance is removed first).

For zeros at 90°, Procedure B fails and Procedure C gives a synthesis with 6 elements and 2 terminations.

To illustrate these ideas, consider

$$Z = \frac{s^3 + 9s^2 + 27s + 3}{s^3 + \frac{11}{3}s^2 + 9s + 25}$$

$$\text{num EvZ} = -s^6 - 3s^4 - 7s^2 + 75 = (3 - s^2)(s^4 + 6s^2 + 25)$$

Even Z has a pair of real zeros at $\pm\sqrt{3}$ and a quadruplet of complex zeros at $\pm 1 \pm j2$. Choosing $a = 1 + j2$ and $b = 1 - j2$ makes ζ_2 four less in rank than Z. Then

$$Z_a = \frac{6}{5}(2 + j1), \quad \frac{\zeta_{1b}}{Z_a} = \frac{3}{5}$$

$$V_{12} = \frac{5 + \frac{24}{5}s + \frac{5}{3}s^2}{5 + \frac{10}{9}s + \frac{3}{5}s^2}, \quad \text{num EvV}_{12} = s^4 + 6s^2 + 25$$

$$\zeta_2 = \frac{\frac{3}{5}s + \frac{3}{5}}{\frac{5}{3}s + 5}, \quad \text{num Ev}\zeta_2 = 3 - s^2$$

Since the $-s^2$ coefficient of num EvV_{12} is of improper sign, Procedure A is not applicable. Procedure C is always applicable. Procedure B is applicable if (35) is satisfied. It is, with the equal sign, in this case. Thus V_{12} is split according to (30).

$$V_{12} = \frac{5 + \frac{3}{5}s^2}{5 + \frac{10}{9}s + \frac{3}{5}s^2} + \frac{\frac{24}{5}s + \frac{16}{15}s^2}{5 + \frac{10}{9}s + \frac{3}{5}s^2} = V_{12a} + V_{12b}$$

The two parts, with their terminations, are expanded to give

$$V_{12a}\zeta_2 = \frac{1}{\frac{1}{\zeta_2} + \frac{1}{\frac{9}{2s} + \frac{27s}{50}}}\,, \quad V_{12b}\zeta_2 = \frac{1}{\frac{25}{24s} + \frac{1}{\frac{8}{s} + \frac{16\zeta_2}{9}}}$$

The network is shown in Figure 7. It has 10 elements (three for each termination). A synthesis by Procedure C would require 17 elements. As an alternative procedure, the bilinear operator in (9) could first be utilized to realize the real even part zeros of Z. The network of Figure 3 would result. Each termination would be two less in rank than Z. Next a biquadratic operator could be formed from each termination and synthesized using Procedure B. The resulting network would be more expensive in elements (requiring 14) but does illustrate the flexibility of the operator approach in permitting the user to choose between several different forms of final network topology.

V BICUBIC OPERATOR SYNTHESIS

The bicubic operator is given by (16). Its even part, using (6), is

$$EvV_{123} = \frac{(a^2 - s^2)(b^2 - s^2)(c^2 - s^2)}{(P_1^2 - Q_1^2)} \tag{42}$$

where P_1 and Q_1 are the even and odd parts, respectively, of the denominator of V_{123}.

A. Half-N Split

Let num EvV_{123} in (42) be expanded to give

$$\begin{aligned} \text{num } EvV_{123} = {} & a^2b^2c^2 - (a^2b^2 + b^2c^2 + c^2a^2)s^2 \\ & + (a^2 + b^2 + c^2)s^4 - s^6 \end{aligned} \tag{43}$$

The coefficients in parentheses in (43) must be positive if a Half-N split is to succeed. Let c be real and a and b be given by (32). Then $a^2b^2c^2$ is certainly positive. The $-s^2$ and s^4 coefficients become, respectively,

$$\left.\begin{aligned} a^2b^2 + b^2c^2 + c^2a^2 &= 2c^2(x^2 - y^2) + (x^2 + y^2)^2 \\ a^2 + b^2 + c^2 &= c^2 + 2(x^2 - y^2) \end{aligned}\right\} \quad (44)$$

Thus c must lie within the range

$$2(y^2 - x^2) \leq c^2 \leq \frac{(x^2 + y^2)^2}{2(y^2 - x^2)} \quad (45)$$

For even part zeros no further than 60° from the real axis ($y \leq \sqrt{3}\,x$), it is always possible to find a positive real c satisfying (45). For roots ≤ 45°, any positive real c will do, whereas for roots at 60°, only $c = 2x$ is satisfactory. For roots between 45° and 60°, there is a finite range of permitted values of c.

Assume (45) is satisfied and write EvV_{123} as the sum of 2 terms.

$$\left.\begin{aligned} EvV_{123} &= \frac{a^2b^2c^2 - (a^2b^2 + b^2c^2 + c^2a^2)s^2}{P_1^2 - Q_1^2} + \frac{(a^2 + b^2 + c^2)s^4 - s^6}{P_1^2 - Q_1^2} \\ &= EvV_{123a} + EvV_{123b} \end{aligned}\right\} \quad (46)$$

Two reactive elements may be removed from each of the parts of (46), leaving in each case, a bilinear remainder operator operating on ζ_3. The synthesis, requiring four elements and two terminations, appears in Figure 8a, where the bilinear operators are denoted by V_{r1} and V_{r2}. Assume that a and b are chosen to realize a quadruplet of complex even part zeros of Z. Assume further that c is chosen to realize a pair of real even part zeros of Z and that this choice of c satisfies (45). Then ζ_3 is 6 less in rank and $V_{r1}\zeta_3$ and $V_{r2}\zeta_3$ are 4 less in rank than Z. As pointed out following (45), it is always possible to choose such a c if the complex quadruplet of even part zeros lies within 45° of the real axis and it may be possible if they lie between 45° and 60°.

When the roots lie between 45° and 60° but c cannot be chosen to realize a pair of real even part zeros, it may be chosen so that one of the equal signs in (45) applies. Such a choice eliminates either the s^2 or s^4 term from (46). Assume that s^2 term is made to vanish. Then V_{123a} may be synthesized by three reactive elements and a resistance. Two reactive elements may be removed from V_{123b}, leaving again a bilinear remainder operator which may be combined with ζ_3. The entire synthesis, requiring five elements and two terminations, appears in Figure 8b. The upper termination is four less in rank than Z, the lower two less.

If the complex roots lie at 60°, choosing $c = 2x$ causes both equal signs in (45) to apply. Now both the s^2 and s^4 terms vanish from (46) and V_{123a} and V_{123b} may each be synthesized with three reactive elements and a resistance.

B. Even-Odd Split

A synthesis similar to Procedure C for the biquadratic operator can also be developed for the bicubic operator. Let EvV_{123} in (42) be separated to give

$$EvV_{123} = \frac{[(a+b+c)s^2+abc]^2}{P_1^2 - Q_1^2} = \frac{-s[s^2+ab+bc+ac]^2}{P_1^2 - Q_1^2} = EvV_{123a} + EvV_{123b} \qquad (47)$$

A shunt capacitor may be removed from V_{123a} whereas V_{123b} has a removable shunt inductor. The remaining biquadratic operators are minimum functions. Each may be synthesized without coupling by the method applied to $V_{12a}\zeta_2$ in Section IIIC. The network appears in Figure 9. The procedure is unrestricted as to complex even-part zero locations and, if a, b and c are utilized for zero cancellation, a rank reduction of six is obtained.

VI BICUBIC OPERATOR SUMMARY

The results of the last section may be summarized by an example. Consider a driving-point impedance given by

$$Z = \frac{s^3 + 9s^2 + 57s + 25}{s^3 + 8s^2 + 5s + 4}$$

with an even part numerator of

$$\text{numEvZ} = -s^6 + 10s^4 - 49s^2 + 100 = (4 - s^2)(25 - 6s^2 + s^4)$$

Choosing

$$a = 2 + j1, \quad b = 2 - j1, \quad c = 2$$

will make ζ_3 a pure resistance (six less in rank than Z).

A Half-N split is more economical in elements than an Even-Odd split. The former requires that condition (45) be satisfied. It is in this case.

Proceeding with the Half-N synthesis,

$$Z_a = 3 - j1, \quad Z_b = 3 + j1, \quad \frac{\zeta_{1b}}{Z_a} = 5$$

$$V_{12} = \frac{5 + 40s + \frac{1}{5}s^2}{5 + \frac{4}{5}s + 5s^2}, \quad \zeta_2 = \frac{5s + 5}{\frac{s}{5} + \frac{4}{5}}, \quad \zeta_{2c} = \frac{25}{2}$$

$$V_3 = \frac{2 + \frac{25}{2}s}{2 + \frac{2}{25}s}, \quad \zeta_3 = \frac{4}{25}$$

$$V_{123} = V_{12}V_3 = \frac{10 + \frac{285}{2}s + \frac{18}{5}s^2 + \frac{5}{2}s^3}{10 + 2s + 20s^2 + \frac{2}{5}s^3}$$

$$EvV_{123} = \frac{100 - 49s^2}{P_1^2 - Q_1^2} + \frac{10s^4 - s^6}{P_1^2 - Q_1^2} = EvV_{123a} + EvV_{123b}$$

V_{123a} and V_{123b} are found from their even parts to be

$$V_{123a} = \frac{10 + \frac{415}{3} s + \frac{83}{30}s^2}{10 + 2s + 20s^2 + \frac{2}{5}s^3}, \quad V_{123b} = \frac{\frac{25}{6}s + \frac{5}{6}s^2 + \frac{5}{2}s^3}{10 + 2s + 20s^2 + \frac{2}{5}s^3}$$

Two reactive elements may be removed from each.

$$V_{123a} = \cfrac{1}{\frac{12}{83}s + \cfrac{1}{\frac{6889}{1380}s + \cfrac{\frac{4067}{46}s + 10}{\frac{46}{83}s + 10}}}, \quad V_{123b} = \cfrac{1}{\frac{12}{5s} + \cfrac{1}{\frac{25}{84s} + \cfrac{\frac{5}{2}s + \frac{5}{7}}{\frac{2}{5}s + 14}}}$$

Thus

$$V_{r1}\zeta_3 = \frac{10\zeta_3 + \frac{4067}{46}s}{10 + \frac{46}{83}s\zeta_3} = \frac{4 + \frac{20335}{92} s}{25 + \frac{92}{415} s}$$

$$V_{r2}\zeta_3 = \frac{\frac{5}{7}\zeta_3 + \frac{5}{2}s}{14 + \frac{2}{5}s\zeta_3} = \frac{\frac{2}{7} + \frac{25}{4}s}{35 + \frac{4}{25}s}$$

and

$$\text{num } EvV_{r1}\zeta_3 = 100 - 49s^2, \quad \text{num } EvV_{r2}\zeta_3 = 10 - s^2$$

The network appears in Figure 10.

VII SYNTHESIS BY PARTIAL MINIMUM RESISTANCE REMOVAL

Recall that the s^2 term of EvV_{12} in (22) vanishes when the even part zeros are at 45°. Also recall that the s^2 and s^4 terms of EvV_{123} in (46) vanish when the even part zeros lie at 60° if $c = 2x$. Thus an economy in elements can be achieved if even part zeros occur at either of these two angles.

In general, these two angle locations are unpopulated, but it is often possible to shift a quadruplet of complex even part zeros there by removing a portion of Z's minimum resistance. Such a procedure can easily be programmed on a digital computer through root locus techniques.

The following two examples, using a biquadratic impedance, illustrate the method. No complications arise for impedances of higher rank because one is always dealing with the biquadratic or bicubic operator, rather than with the impedance itself.

1. $$Z = \frac{\frac{3}{2}s^2 + \frac{69}{20}s + \frac{33}{10}}{s^2 + 4s + 5}, \quad \text{num } EvZ = \frac{3}{2}(s^4 - 2s^2 + 11)$$

The even part zeros occur at $s = \pm 1.47 \pm j1.08$. The least number of elements by which Z can be synthesized using the methods of Section III is eight. But if the even part zeros can be shifted to $\pm 45^o$ from the real axis, the synthesis requires only seven elements. The shift is accomplished by removing 1/2 ohm from Z.

$$Z' = Z - \frac{1}{2} = \frac{s^2 + \frac{29}{20}s \; \frac{4}{5}}{s^2 + 4s + 5}, \quad \text{num } EvZ' = s^4 + 4$$

Application of Procedure A of Section III yields the network of Figure 11.

2. $$Z = \frac{\frac{6}{5}s^2 + \frac{41}{20}s + \frac{21}{5}}{s^2 + 4s + 5}, \quad \text{num } EvZ = \frac{6}{5}(s^4 + \frac{5}{3}s^2 + \frac{35}{2})$$

The even part zeros lie at $s = \pm 1.30 \pm j1.58$. As resistance is removed the even part zeros of the remaining impedance move away from the real axis. Hence they never reach $\pm 45^o$. However they can be shifted to $\pm 60^o$ if 1/5 ohm is removed.

$$Z' = Z - \frac{1}{5} = \frac{s^2 + \frac{21}{20}s + \frac{16}{5}}{s^2 + 4s + 5}, \quad \text{num } EvZ' = s^4 + 4s^2 + 16$$

Now Procedure A of Section V may be applied to yield a synthesis with nine elements. The values are

$$a = 1 + j\sqrt{3}, \quad b = 1 - j\sqrt{3}, \quad c = 2x = 2$$

$$Z_a = \frac{9+j\sqrt{3}}{20}, \quad \zeta_{1b} = \frac{4}{5}Z_a, \quad \zeta_2 = \frac{16}{25} = \zeta_3$$

$$V_{12} = \frac{4 + \frac{21}{25}s + \frac{5}{4}s^2}{4 + 5s + \frac{4}{5}s^2}, \quad V_3 = \frac{2 + \frac{16}{25}s}{2 + \frac{25}{16}s}$$

$$V_{123} = \frac{8 + \frac{106}{25}s + \frac{61}{16}s^2 + \frac{4}{5}s^3}{8 + \frac{65}{4}s + \frac{24}{5}s^2 + \frac{5}{4}s^3}, \quad \text{num } EvV_{123} = 64 - s^6$$

$$V_{123a} = \frac{8 + \frac{1152}{452}s + \frac{12}{17}s^2}{8 + \frac{65}{4}s + \frac{24}{5}s^2 + \frac{5}{4}s^3}, \quad V_{123b} = \frac{\frac{26}{17}s + \frac{845}{272}s^2 + \frac{4}{5}s^3}{8 + \frac{65}{4}s + \frac{24}{5}s^2 + \frac{5}{4}s^3}$$

V_{123a} and V_{123b} may each be synthesized with three reactive elements and a resistive termination. The network is shown in Figure 12.

Footnotes

* This paper is based on work supported by the National Science Foundation under grant NSF G-23426.

1. D. Hazony, "Zero Cancellation Synthesis Using Impedance Operators," IRE Trans. on Circuit Theory, vol. CT-8, pp. 114-120, June, 1961.
2. J.B. Murdoch and D. Hazony, "Cascade Driving-Point Impedance Synthesis by Removal of Sections Containing Arbitrary Constants," IRE Trans. on Circuit Theory, vol. CT-9, pp. 56-61, March, 1962.
3. Sum of degrees of numerator and denominator.
4. Z is assumed minimum reactive and susceptive. The operator approach also applies to admittances.
5. J.B. Murdoch, "Driving-Point Impedance Synthesis Using Impedance Operators," Report No. 27, AF19(604)-3887, Case Institute of Technology, August, 1961.
6. The proof is obtained by direct expansion.
7. ibid 1.
8. P.I. Richards, "A Special Class of Functions with Positive Real Part in Half Plane," Duke Math J., vol. 14, pp. 777-788. September, 1947.
9. ibid 1.
10. ibid 5.
11. E.A. Guillemin, Passive Network Synthesis, John Wiley and Sons, New York, 1957.
12. ibid 11.
13. E.S. Kuh, "Special Synthesis Techniques for Driving Point Impedance Functions," IRE Trans. on Circuit Theory, vol. CT-2, pp. 302-308, December, 1955.
14. ibid 11.
15. ibid 1. Hazony's method removes $V_{12}/_s$ minimum resistance and utilizes a surplus factor rather than an additional operator.
16. ibid 1. The proof of (41) is similar and will not be repeated here.

FIGURES

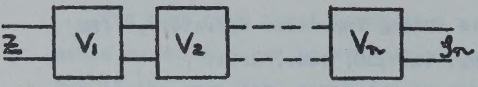


FIG.1- CASCADE REPRESENTAION OF Z

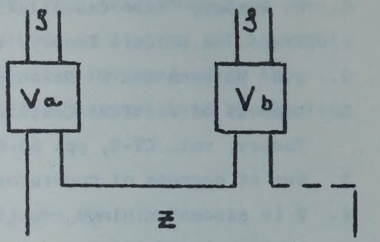


FIG.2-DISTRIBUTED REPRESENTAION OF Z

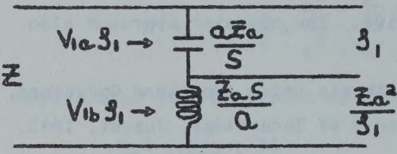

V1a g1 →
aZa/S
g1
Z
Za S/a
V1b g1 →
Za²/g1

FIG.3-BILINEAR OPERATOR SYNTHESIS

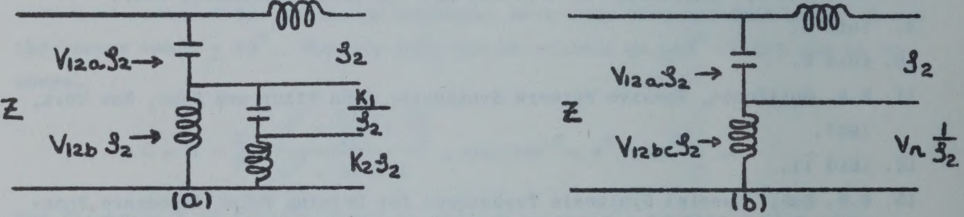

V12a g2 →
g2
Z
K1/g2
V12b g2 →
K2 g2
(a)
V12a g2 →
g2
Z
V12bc g2 →
Vn 1/g2
(b)

FIG 4-BIQUADRATIC OPERATOR SYNTHESIS A

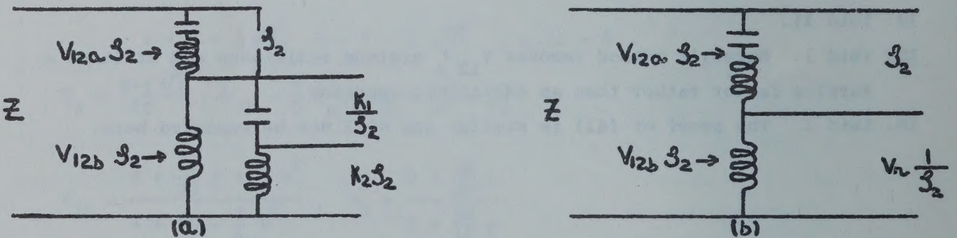

V12a g2 →
g2
Z
K1/g2
V12b g2 →
K2 g2
(a)
V12a g2 →
g2
Z
V12b g2 →
Vn 1/g2
(b)

FIG.5-BIQUADRATIC OPERATOR SYNTHESIS B

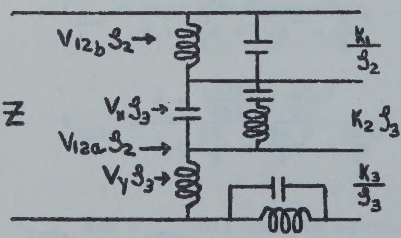

V12b g2 →
K1/g2
Z
Vx g3 →
K2 g3
V12a g2 →
Vy g3 →
K3/g3

FIG.6-BIQUADRATIC OPERATOR SYNTHESIS C

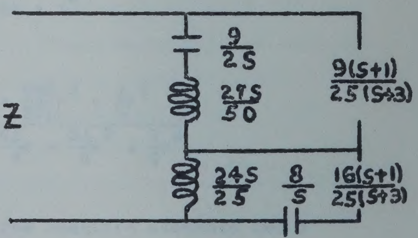

9/2.5S
2.7S/5.0
9(S+1)/2.5(S+3)
Z
2.4S/2.5
8/S
16(S+1)/2.5(S+3)

FIG.7— $Z = \frac{S^3+9S^2+27S+3}{S^3+\frac{11}{2}S^2+9S+2.5}$

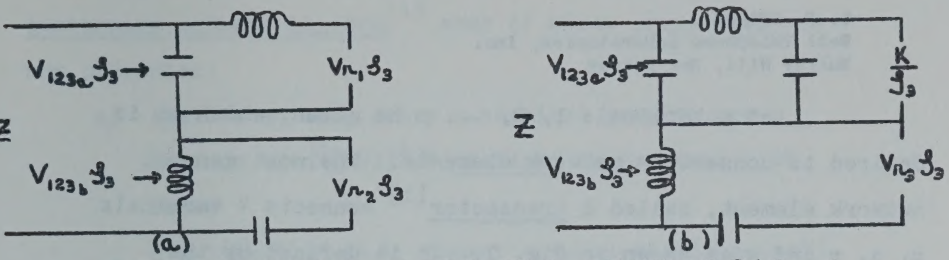

V123a g3 →
Vn1 g3
Z
V123b g3 →
Vn2 g3
(a)
V123a g3 →
K/g3
Z
V123b g3 →
Vn3 g3
(b)

FIG.8—BICUBIC OPERATOR SYNTHESIS A

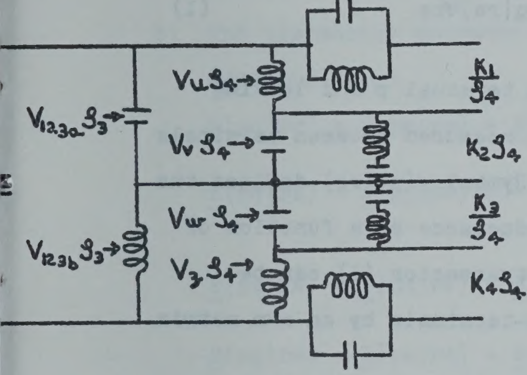

Vu g4 →
K1/g4
V123a g3 →
Vv g4 →
K2 g4
Vw g4 →
K3/g4
V123b g3 →
Vy g4 →
K4 g4

FIG.9—BICUBIC OPERATOR SYNTHESIS B

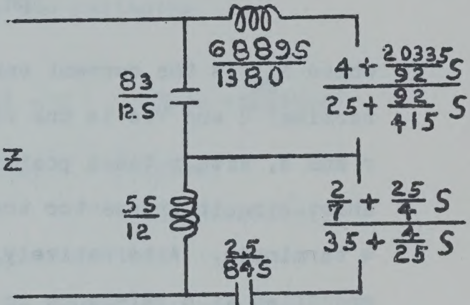

68895/1380
83/12.5
4 + (20335/92) s / 25 + (92/415) s
Z
5s/12
2/7 + (25/4) s / 35 + (4/25) s
2.5/84s

FIG.10

$$Z = \frac{s^3 + 9s^2 + 57s + 25}{s^3 + 8s^2 + 5s + 4}$$

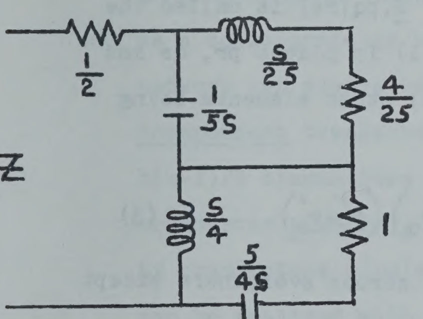

1/2
s/25
4/25
1/5s
Z
s/4
1
5/4s

FIG.11

$$Z = \frac{\frac{3}{2}s^2 + \frac{69}{20}s + \frac{33}{10}}{s^2 + 4s + 5}$$

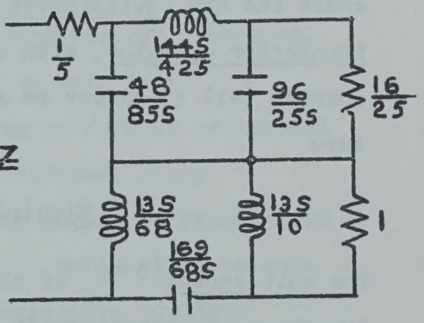

1/5
144s/425
48/85s
96/25s
16/25
Z
13s/68
13s/10
1
169/68s

FIG.12

$$Z = \frac{\frac{6}{5}s^2 + \frac{41}{20}s + \frac{21}{5}}{s^2 + 4s + 5}$$

A QUATERNARY NETWORK ALGEBRA

G. E. SHARPE
Bell Telephone Laboratories, Inc.
Murray Hill, New Jersey

Let n-terminals 1, 2, ... n be given, which it is desired to connect by network elements. The most general network element, called a transactor[1] connects 4 terminals p, q, r and s as shown in Fig. 1. It is defined by the current voltage equation

$$i_{pq} = y(pq|rs)Vrs \qquad (1)$$

where i_{pq} is the current entering terminal p and leaving terminal q and Vrs is the voltage measured between terminals r and s, with r taken positive. Symbol y(pq|rs) defines the short-circuit transactor transconductance as a function of 4 terminals. Alternatively, the transactor (1) can be specified with reference to the n-terminals by an n×n matrix operator $\underset{\sim}{t}(pq|rs)$

$$\underset{\sim}{t}(pq|rs) = y(pq|rs)\underset{\sim}{E}(pq|rs) \qquad (2)$$

where the n×n equicofactor matrix[2] $\underset{\sim}{E}(pq|rs)$ is called the transactor operator, with entries (+1) in places pr, qs and entries (-1) in places ps and qr, all other elements being zero

$$\underset{\sim}{E}(pq|rs) = \left(\underset{\sim}{e}_p - \underset{\sim}{e}_q\right)\left(\underset{\sim}{e}'_r - \underset{\sim}{e}'_s\right) \qquad (3)$$

The unit vector[3] $\underset{\sim}{e}_p$ of order n has zeroes everywhere except for the p-th element which is unity

$$\underset{\sim}{e}_p = (0,0\ldots1,\ldots0)' \qquad (4)$$

and where the superscript prime implies matrix transpose. As a consequent of (3) both $\underset{\sim}{t}(pq|rs)$ and $\underset{\sim}{E}(pq|rs)$ are simple products of rank 1. These operators form the basis of a new quaternary network algebra,[4] some of whose properties are now presented:

a) Associated with every n-terminal network are a maximum of $n^2(n-1)^2/4$ distinct transactors. Of these
$n(n-1)(n-2)(n-3)/4$ connect 4 terminals $[\underset{\sim}{t}(pq|rs)]$,
$n(n-1)(n-2)$ connect 3 terminals $[\underset{\sim}{t}(pq|rq)]$, and
$n(n-1)/2$ connect 2 terminals $[\underset{\sim}{t}(pq|pq)]$.

b) The transactor operators obey the following addition laws:

$$\underset{\sim}{E}(pq|rr) = \underset{\sim}{E}(pp|rs) = \underset{\sim}{E}(pp|pp) = \underset{\sim}{0} \qquad \text{(Zero operator)}$$

$$\underset{\sim}{E}(pq|rs) = \underset{\sim}{E}(qp|sr) = -\underset{\sim}{E}(pq|sr) = -\underset{\sim}{E}(qp|rs) \qquad \text{(Sign reversal)}$$

$$\begin{aligned} \underset{\sim}{E}(pq|rs) &= \underset{\sim}{E}(pq|rv) - \underset{\sim}{E}(pq|sv) \\ \underset{\sim}{E}(pq|rs) &= \underset{\sim}{E}(pu|rs) - \underset{\sim}{E}(qu|rs) \qquad \text{(Addition)} \qquad (5) \\ \underset{\sim}{E}(pq|rs) &= \underset{\sim}{E}(pu|rv) + \underset{\sim}{E}(qu|sv) - \underset{\sim}{E}(pu|sv) - \underset{\sim}{E}(qu|rv) \end{aligned}$$

with references u and v arbitrary.

c) As a consequence of b) every nonreciprocal n-terminal network can always be realized by only $(n-1)^2$ independent transactors. If the network is reciprocal $n(n-1)/2$ transactors suffice. Further, every transactor $\underset{\sim}{t}(pq|rs)$, $(pq \neq rs)$ can always be represented by transactors $\underset{\sim}{t}(pu|su)$, $(ps \neq u)$. Hence every network can be realized with $(n-1)^2$ transactors $\underset{\sim}{t}(pu|su)$ and $\underset{\sim}{t}(pq|pq)$, which connect only 3 and 2 terminals respectively. If the admittance matrix of the n-terminal network is $\underset{\sim}{Y}$, this may be stated for the optimum case as

e) As a consequence of d) all operators $\underset{\sim}{E}(pu|su)$ equal their square and are called *idempotent*, as are also all weighted operators $\frac{1}{\sqrt{2}}\underset{\sim}{E}(pq|pq)$. Further, it is possible to define a *general network inverse*[5] Z, where

$$\underset{\sim}{YZ} = \underset{\sim}{ZY} = \underset{\sim}{C} \qquad (7)$$

$\underset{\sim}{Z}$ is an n×n equicofactor matrix of impedance coefficients and $\underset{\sim}{C}$ is the n×n *centering matrix*

$$\underset{\sim}{C} = \underset{\sim}{I} - \frac{\underset{\sim}{e}\underset{\sim}{e}'}{n} \qquad (8)$$

where $\underset{\sim}{I}$ is the unit matrix and $\underset{\sim}{e} = (1,1,1,\ldots 1)'$ is the sum vector and both are of order n.

I have now presented some elements of the new quaternary network algebra and have given a few examples of its application. It is hoped this talk will stimulate others to apply this useful new tool to the further analysis and synthesis of active networks.

REFERENCES

† Bell Telephone Laboratories, Inc., Murray Hill, New Jersey.

1. G. E. Sharpe, "Axioms on Transactors", IRE, PGCT, Vol. CT-5, pp. 189-196, September 1958.

2. G. E. Sharpe and B. Spain, "On the Solution of Networks by Means of the Equicofactor Matrix," IRE, PGCT, Vol. CT-7, No. 3, pp. 230-238, September 1960.

3. E. Bodewig, "Matrix Calculus", North Holland Publishing Company, 1959.

4. G. E. Sharpe, "The Realization and Measurement of n-Terminal Active Networks", submitted for publication, IRE, PGCT.

5. G. E. Sharpe and G. P. H. Styan, "Circuit Duality and the General Network Inverse", to be submitted for publication, IRE, PGCT.

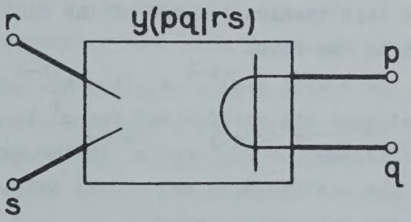


FIG. 1
VOLTAGE-CURRENT TRANSACTOR-VCT, DEGREE 4

ON THE USE OFTHE PASCAL TRIANGLE WITH NETWORK FUNCTIONS

K. A. PULLEN
Ballistic Research Laboratories
Aberdeen Proving Ground, Maryland

The Pascal Triangle

The Pascal triangle, Table I, is a triangular array of coefficients or numbers which have been extracted from the binomial expansion, $(1 + x)^n$. In addition to their derivation from the binomial expansion, they may also be derived in other ways. For example, each number in Table I may be obtained by adding the numbers adjacent and above them. (Numbers formed in this manner are sometimes called Fibonacci numbers.) This array as a result can be generated directly from the diagonal ones which form the limiting values for the triangle.

TABLE I
The Pascal Triangle

m = 0	1						
m = 1	1	1					
m = 2	1	2	1				
m = 3	1	3	3	1			
m = 4	1	4	6	4	1		
5	1	5	10	10	5	1	
	1	6	15	20	15	6	1

The numbers generate in this fashion because of the fact the generating equation may be written in the form:

$$y = (1 + x)^n = (1 + x)(1 + x)^{n-1} = (1 + x)^{n-1} + x(1 + x)^{n-1} \quad (1)$$

From this, it is evident that the coefficient for x^r in $(1 + x)^n$ must be the sums of the coefficients of x^{r-1} and x^r in the expansion of $(1 + x)^{n-1}$. These are the coefficients above left and above right of the desired coefficient in the triangular array.

Because of its nature, the Pascal triangle must include all integers and, as a result, it is possibly not surprising that it should include the coefficients for many common mathematical functions used in network theory. What possibly is interesting, however, is the fact that complete sets of these numbers can be found in the array, and these sets obey precise relations with respect to one-another.

The numbers in the binomial expansion are normally specified in terms of quotients of factorials. The coefficients for x^r in the expansion of $(1 + x)^n$ are typically written in the form:

$$\binom{n}{r} = (n!) / (r!)(n-r)! \quad (2)$$

This relation is adequate for positive integral values of n, but under other conditions requires modification. The factorials are replaced by gamma functions of (n+1) and (n-r+1) for non-integral values of n, and they are replaced by $(-1)^r\binom{n+r-1}{r}$ for negative integral values of n. (The factoring of the minus sign from the coefficient makes the change in terminology necessary.)

Coefficient Relations

The form taken by the coefficients for the binomial with a negative integral exponent indicates that possibly additional groups of coefficients might be of use. For the general case, for example, the coefficients might take the form:

$$\binom{m+ku}{ju} = (m+ku)!\ /\ (ju)!\ (m+[k-j]u)! \qquad (3)$$

Typical values of j and k are small integers; for example, one and two. The limiting value is specified for k = 2j, since this coordinate relation specifies the line of maximum values, namely, vertically downward. For k = 3j, the results are the same as for 2k = 3j, and as a result, no new data are developed.

In Equation 3, a self-limiting sequence results as long as the values of k and j satisfy the relation:

$$k < j \qquad (4)$$

For a value of k equal to or greater than that of j, an infinite sequence of numbers results. The self-limiting sequences prove to be useful with common polynomials, particularly the sets of orthogonal polynomials, and the infinite sequences with such functions as Bessel functions.

The generating differential equations for the various coefficient sequences have been established, and most of these have proven to be forms of the hypergeometric differential equation. All of them satisfy hypergeometric-type dquations having two regular singularities (typically either zero and plus (or minus) one, or zero and minus four) and an irregular singularity resulting from the confluence of two singularities at infinity. The exponent differences on the regular singularities typically are 1/2, and the exponents at infinity are related to the order of the polynomial or function. (One exponent must be negative for the development of a polynomial.)

The Polynomials for Continued-Fractions

The polynomials which are approximants for the continued fraction:

$$\begin{aligned} U(x) &= Y + 1 \rfloor\!\lceil Z + 1 \rfloor\!\lceil Y + 1 \rfloor\!\lceil Z + 1 \rfloor\!\lceil Y + \dots && (5a)\\ &= Y + Y \rfloor\!\lceil U + 1 \rfloor\!\lceil 1 + 1 \rfloor\!\lceil U + 1 \rfloor\!\lceil 1 + \dots && (5b)\end{aligned}$$

where U = YZ, are of considerable importance with ladder networks. The two polynomials which are the n'th order approximants for this fraction, and which are the equations for the poles and the zeros for the simple R-C network, are defined by the equations:

$$F_n(U) = \sum_{u=0}^{m} \binom{m+u}{2u} U^u \qquad (6a)$$

$$Y_n(U) = \sum_{u=0}^{m-1} \binom{m+u}{2u+1} U^u \qquad (6b)$$

Clearly, the coefficients for these polynomials are readily extracted from Eq. 3, by taking $k = 1$ and $j = 2$. As long as m is an integer, the value of the coefficient may be read directly from the Pascal triangle.

The definition of the reqursion equations for the continued fraction may be used in developing these polynomials:

$$A_{2j} = b_{2j-1}A_{2j-1} + A_{2j-2} \qquad B_{2j} = b_{2j-1}B_{2j-1} + B_{2j-2} \qquad (7a)$$

$$A_{2j+1} = b_{2j}A_{2j} + A_{2j-1} \qquad B_{2j+1} = b_{2j}B_{2j} + B_{2j-1} \qquad (7b)$$

When these equations are applied either to (5a) or (5b), the recursion relations become:

$$\begin{aligned} A_{2j+1} &= Y\,A_{2j} + A_{2j-1} \\ A_{2j} &= Z\,A_{2j-1} + A_{2j-2} \end{aligned} \qquad (8)$$

Combining gives the recursion relations in the two forms:

$$A_{2j+1} = (U + 2)\,A_{2j-1} - A_{2j-3} \qquad (9)$$

or

$$F_{n+1} = (U + 2)\,F_n - F_{n-1} \qquad (10)$$

The Y_n function satisfies an identical recursion relation, and the following relations may be established between the two functions:

$$\begin{aligned} F_{n+1} &= (U + 1)\,F_n \underset{x}{+} U\,Y_n \\ Y_n &= Y_{n-1} + F_{n-1} \end{aligned} \qquad (11)$$

These subordinate relations may be called sub-recursion relations when they occur in connection with such polynomial systems.

The F and the Y polynomials may also be generated by the following matrix procedure:

$$\begin{pmatrix} (1+U) & Z \\ Y & 1 \end{pmatrix}^n = \begin{pmatrix} F_n & ZY_n \\ YY_n & F_{n-1} \end{pmatrix} \qquad (12)$$

In a ladder network, as successive sections are added, the addition of the n'th section increases the number of sections from n-1 to n, and also post-multiplies the (n-1) order matrix by the basic section matrix. It may be shown by induction that the above relation is true through the use of the sub-recursion relations of Eq. 11.

Lattice Network Polynomials

The interesting resulte obtained with the ladder network causes one to ask the question, "How about the lattice network?" The lattice network also has associated with it some polynomials. These polynomials take quite a different form from those for the ladder network, but once again, their coefficients may be derived directly from the coefficients in the Pascal triangle. The polynomials are specified in terms of the summations:

$$G_n(U) = \sum_{r=0}^{n} \binom{2n}{2r} U^r \tag{13}$$

$$Z_n(U) = \sum_{r=0}^{n} \binom{2n}{2r+1} U^r \tag{14}$$

Typically, for example, the coefficients for G_3 are 1, 15, 15, and 1, extracted from the series, 1, 6, 15, 20, 15, 6, and 1; and the coefficients for Z_3 are 6, 20, and 6.

The recursion relation for the lattice polynomials takes the general form:

$$G_{n+1} = 2(1 + U)\, G_n - (1 - U)^2\, G_{n-1} \tag{15}$$

In this equation, the G functions may be replaced by the Z functions to obtain the corresponding relation for the Z functions.

The forms of Eqs. 10 and 15 are of interest inasmuch as they give some clues to the form of the associated network. The presence of a non-unity coefficient on the G_{n-1} term indicates that there is more than one path through the network section between input and output. In addition, the coefficient of the middle term (which is the trace of the A matrix) can be used to specify the open-circuit and short-circuit characteristics of the individual section. The general form for Eq. 15 is:

$$M_{n+1} = (A_i + D_i)\, M_n + (UB_iC_i - A_iD_i)\, M_{n-1} \tag{16}$$

where A_i, B_i, C_i and D_i are elements of the general transmission matrix, and M_n is an n'th order polynomial or function. The general transmission matrix takes the form:

$$[A] = \begin{pmatrix} A_i & B_i \\ C_i & D_i \end{pmatrix} \tag{17}$$

Since the A_i is the forward voltage ratio, and the D_i is the forward current ratio, the coefficient of the M_n term is numerically proportional to the sum of the number of network trees for the voltage-gain configuration and the current-gain configuration.

The coefficient of the M_{n-1} term may be converted into the form:

$$(UB_iC_i - A_iD_i) = B_iC_i(U - A_iD_i/B_iC_i) = B_iC_i(U - Y_IZ_I) \tag{18}$$

where Y_I and Z_I are the input admittance (short-circuit) and the input impedance (open-circuit) respectively. This reduces to unity for all cascade network sections having only one through-path, and takes a form which specifies any network zeroes introduced by parallel paths which are present in multiple-path sections. (The coefficient for the twin-tee network takes the form $(1 + U^2)^2$, for example.)

The Generation of Legendre Coefficients

Several sets of polynomials used by both mathematicians and engineers can also be generated directly with the aid of the Pascal triangle. With these polynomials, only a few can be read directly from the triangle without the use of binomial scaling coefficients. For example, with the Legendre polynomials, when expressed in terms of powers of "x", the coefficients for the symmetrical form are:

$$P_n(x) = (1/2)^{n-1} \sum_{r=0}^{[n/2]} \left[(-1)^r \binom{2n-2r-1}{n-r} \binom{n-r}{r} x^{n-2r}\right] \qquad (19)$$

The two binomial coefficients used in this series of equations must be multiplied by two additional factors, one involving a power of negative unity, and the other a fixed power of one half.

The Legendre polynomials may also be expanded in a non-symmetric form by making either of the substitutions:

$$\begin{aligned} x &= 2y + 1 \qquad x = [-1, 1] \qquad y = [-1, 0] \\ x &= 2y - 1 \qquad x = [-1, 1] \qquad y = [0, 1] \end{aligned} \qquad (20)$$

In the second equation, if the substitution $y = z^2$ is made, then the result is an algebraic equation similar to the equation:

$$\cos\theta = \sqrt{(1 + \cos 2\theta)/2} \qquad (21)$$

where $x = \cos 2\theta$ and $z = \cos\theta$. The importance of this relation is particularly shown with the Chebychev non-symmetric polynomials.

The nonsymmetric Legendre polynomials may also be expressed in terms of the coefficients of the Pascal triangle:

$$\begin{aligned} P_n(y) &= \sum_{r=0}^{n} \binom{n+r}{2r} \binom{2r}{r} y^r \qquad [-1, 0] \\ R_n(y) &= \sum_{r=0}^{n} \binom{n+r}{2r+1} \binom{2r+1}{r} y^r \qquad [-1, 0] \end{aligned} \qquad (22)$$

where the P_n are the functions obtained by making the transformation, and the R_n are a set of associated half-angle functions.

The coefficients for Legendre polynomials in the multi-angle form (where the expansion is in terms of $\cos n\theta$ instead of $\cos^n \theta$) as well as the polynomial form may be expressed in terms of the Pascal coefficients. Two separate cases must be considered, even-order, and odd-order polynomials. The defining equations are:

$$P_{2n}(\theta) = (1/2)^{4n} \sum_{r=0}^{n} \binom{2n-2r}{n-r} \binom{2n+2r-1}{n+r-1} \cos 2r\theta \, (1 + \delta_{0r})^{-1}$$
$$P_{2n+1}(\theta) = (1/2)^{4n+2} \sum_{r=0}^{n} \binom{2n-2r-2}{n-r-1} \binom{2n+2r-1}{n+r-1} \cos (2r+1)\theta \qquad (23)$$

These coefficients are somewhat complex, but they can be evaluated more easily from the Pascal triangle than from the general defining equations.

Chebychev Polynomials

The Chebychev polynomials are algebraic expressions for the multi-angle trigonometric relations, in accordance with the defining relation:

$$T_n(x) = \cos n(\arccos x) \qquad (24)$$

As a consequence of this relation, the value of x^n may be expressed in terms of trigonometric multi-angle relations and the Pascal coefficients in the form:[α]

$$x^n = \overset{[n/2]}{\underset{r=0}{}} \binom{n}{r} \cos[(n-2r)\theta] \, (1 + {}_{n/2,r})^{-1} \qquad (25)$$

[α] Hamming, Numerical Methods for Scientists and Engineers, p. 255

The coefficients for the polynomials may be expressed in terms of Pascal coefficients in either of the forms:

$$T_{2m}(x) = \sum_{u=0}^{m} (-1)^u (2)^{2u-1} \left[\binom{m+u-1}{2u-1} + 2\binom{m+u-1}{2u} \right] x^{2u}$$
$$= \sum_{u=1}^{m} 2(-4)^{u-1} (m/u) \binom{m+u-1}{2u-1} x^{2u} + \delta_{,u}\ {}_{0,u}$$
$$T_{2m+1}(x) = \sum_{u=0}^{m} (-1)^u (2)^{2u} \left[\binom{m+u}{2u} + 2\binom{m+u}{2u+1} \right] x^{2u+1} \qquad (26)$$
$$= \overset{m}{\underset{u=0}{}} (-4)^u [(2m+1)/(2u+1)] \binom{m+u}{2u} x^{2u+1}$$

These coefficients are somewhat complex to determine, but they are much more easily handled than those obtained by ordinary procedures. The first few of the polynomials are:

TABLE II

Chebychev Polynomials

$T_0 = 1$ $\qquad$ $T_1 = x$
$T_2 = 1 - 2x^2$ $\qquad$ $T_3 = 3x - 4x^3$
$T_4 = 1 - 8x^2 + 8x^4$ $\qquad$ $T_5 = 5x - 20x^3 + 16x^5$
$T_6 = 1 - 18x^2 + 48x^4 - 32x^6$

and so forth.

The symmetric form of these equations may be transformed into an unsymmetric form by the substitution of $x = 2y + 1$, giving the set:

TABLE III

Nonsymmetric Chebychev Polynomials

$$T_0(y) = 1 \qquad T_1(y) = 2y + 1$$
$$T_2(y) = 8y^2 + 8y + 1 \qquad T_3(y) = 32y^3 + 48y^2 + 18y + 1$$

and so forth. Clearly, this table is constructed from the previous one by the substitution of $(-y)$ for x^2 in the even functions. The nonsymmetric odd functions are half-order functions, and they take the form:

TABLE IV

Associated Nonsymmetric Chebychev Polynomials

$$V_0(y) = 1 \qquad V_1 = 4y + 3$$
$$V_2(y) = 16y^2 + 20y + 5 \qquad V_3 = 64y^3 + 112y^2 + 56y + 7$$
$$V_4(y) = 256y^4 + 576y^3 + 432y^2 + 120y + 9$$
$$V_5(y) = 1024y^5 + 2816y^4 + 2816y^3 + 1232y^2 + 220y + 11$$

The recursion relations for both the symmetrical and the nonsymmetrical forms of these equations, and the sub-recursion relations for the nonsymmetrical form are:

$$T_{n+1} = 2x\,T_n - T_{n-1} \qquad \text{in terms of } x \tag{27}$$

$$T_{n+1} = 2(2y + 1)T_n - T_{n-1} \qquad \text{in terms of } y$$

$$\left.\begin{aligned} T_{n+1} &= T_n + 2y\,V_n \\ V_n &= V_{n-1} + 2T_n \end{aligned}\right\} \qquad \text{in terms of } y \tag{29}$$

An additional useful form of (27) is the equation:

$$T_{n+2}(x) = 2(2x^2 - 1)\,T_n(x) - T_{n-2}(x) \tag{30}$$

The relation between (28) and (30) is self-evident.

Bessel Polynomials

Bessel polynomials are the approximants for the continued fraction representing the arctangent function. They are particularly convenient for use in network synthesis because of the fact they correspond, in the constant-phase problem, to the Butterworth polynomials commonly used for design of monotonically controlled amplitude response. There are two sets of polynomials to consider, the principal ones usually given, and the half-order set corresponding to the Y_n and Z_n functions for the ladder and the lattice networks. The two sets of polynomials are:

TABLE V

Bessel Polynomials

$B_0 = 1$ $\qquad$ C_0 -----

$B_1 = s + 1$ $\qquad$ $C_1 = 1$

$B_2 = 3s^2 + 3s + 1$ $\qquad$ $C_2 = 3s + 2$

$B_3 = 15s^3 + 15s^2 + 6s + 1$ $\qquad$ $C_3 = 15s^2 + 12s + 3$

$B_4 = 105s^4 + 105s^3 + 45s^2 + 10s + 1$

$C_4 = 105s^3 + 90s^2 + 30s + 4$

where the B_n are the integral-order functions, and the C_n the half-integer functions.

The coefficients for these polynomials again can be extracted from the Pascal triangle. In terms of Pascal coefficients, the two sets take the form:

$$B_n(x) = \mathop{}_{r=0}^{n} \binom{n+r}{2r}(2r-1)(2r-3)(2r-5)\ \ldots\ 3.1\ x^r \qquad (31)$$

$$C_n(x) = \mathop{}_{r=0}^{n} \binom{n+r}{2r+1}(2r+1)(2r-1)(2r-3)\ \ldots\ 3.1\ x^r \qquad (32)$$

Recursion relations and sub-recursion relations may be defined in terms of these equations to be:

$$B_{n+1} = (2n+1)\ x\ B_n + B_{n-1} \qquad (33)$$

$$\left.\begin{aligned} B_{n+1} &= B_n + x\ C_{n+1} \\ C_{n+1} &= (2n+1)\ B_n - C_n \end{aligned}\right\} \qquad (34)$$

Summing series may be established for either of these equations if desired.

Bessel Functions

The simplest number sequence in the Pascal triangle which does not terminate is that corresponding to $(1 - x)^{-m}$. The next simplest, the one for which j and k in Eq. 3 have values 1 and 2 respectively, is related to the coefficients in the power expansion for the Bessel function. Expressed in terms of binomial coefficients, the general (real) Bessel

$$H_n(x) = \mathop{}_{r=0}^{\lfloor n/2 \rfloor} (-1)^r \binom{n}{2r}\ (2r-1)(2r-3)(2r-5)\ \ldots\ 3.1\ x^{n-2r} \qquad (36)$$

The recursion relation for the coefficients for the Hermite polynomial is:

$$H_{n+1}(x) = x\ H_n(x) - n\ H_{n-1}(x) \qquad (37)$$

The alternate polynomials associated with H_n prove to be defined by the equation:

$$H^{\alpha}_{\ n} = n\ H_{n-1} \qquad (38)$$

where $H^{\alpha}_{\ n}$ is the associated polynomials. As a result, both the underlined and the non-underlined coefficients in the table represent the same set of functions. Consequently, there are no sub-recursion relations for this set of polynomials.

Laguerre Polynomials

Laguerre polynomials are a set of orthogonal polynomials based on one of the intervals, $[0, \infty]$ or $[-\infty, 0]$. Their generation is based on the use of the weighting factor e^{-x} or e^{x}, respectively. There are two commonly-used forms of these polynomials, the generalized, and the form for which the generalizing factor, α, has the value zero. The general equations for the two forms of the polynomials are:

$$L_k (x) = \underset{u=0}{\overset{k}{}} (-1)^u \binom{k}{u} (\alpha + k)^{(u)} x^{k-u} \tag{39}$$

$$L_k (x) = \underset{u=0}{\overset{k}{}} (-1)^u \binom{k}{u} (k! / u!) x^{k-u} \tag{40}$$

where $(\alpha + k)^{(u)}$ has its usual meaning. The first few polynomials in the reduced set $(\alpha = 0)$ are:

TABLE VIII

Laguerre Polynomials

$L_0 (x) = 1$ $\qquad$ $L_1 (x) = 1 - x$

$L_2 (x) = 2 - 4x + x^2$ $\qquad$ $L_3 (x) = 6 - 18x + 9x^2 - x^3$

$L_4 (x) = 24 - 96x + 72x^2 - 16x^3 + x^4$

$L_5 (x) = 120 - 600x + 600x^2 - 200x^3 + 25x^4 - x^5$

The recursion relation for these polynomials is:

$$L_{k+1} (x) = (2k + 1 - x) L_k (x) - k^2 L_{k-1} (x) \tag{41}$$

These equations may be made applicable in the left-hand plane by a sub-

function has the value:

$$J_n (x) = \sum_{r=0}^{\infty} (-1)^r [(n+2r)!]^{-1} \binom{n+2r}{r} (x/2)^{n+2r} \tag{35}$$

Because of the value of the ratio of j to k, the coefficients fall in vertical columns, the one on the apex being that for J_0, and the one for J_n originating at the coefficient $\binom{n}{0}$.

Hermite Polynomials

Hermite polynomials are a set of orthogonal polynomials based on the interval $[-\infty, \infty]$. The weighting function required in their development is e^{-x^2}. The first few of the polynomials are:

TABLE VI

Hermite Polynomials

$$\begin{array}{llll} H_0 = 1 & & H_1 = x \\ H_2 = x^2 - 1 & & H_3 = x^3 - 3x \\ H_4 = x^4 - 6x^2 + 3 & & H_5 = x^5 - 10x^3 + 15x \\ H_6 = x^6 - 15x^4 + 45x^2 - 15 & & \end{array}$$

The coefficients for these polynomials may be extracted from the Pascal triangle as modified for generating the coefficients for the Bessel polynomials. The modified triangle, with the appropriate coefficients underlined, follows:

TABLE VII

Pascal Array for Hermite Polynomials

$$\begin{array}{ccccccccccccccc}
&&&&&&&\underline{1}&&&&&&& \\
&&&&&&\underline{1}&&1&&&&&& \\
&&&&&\underline{1}&&2&&\underline{1}&&&&& \\
&&&&\underline{1}&&3&&\underline{3}&&3&&&& \\
&&&\underline{1}&&4&&\underline{6}&&12&&\underline{3}&&& \\
&&\underline{1}&&5&&\underline{10}&&30&&\underline{15}&&15&& \\
&\underline{1}&&6&&\underline{15}&&60&&\underline{45}&&90&&15& \\
\underline{1}&&7&&\underline{21}&&105&&\underline{105}&&315&&\underline{105}&&105
\end{array}$$

and so forth. In this instance, the coefficients for H_j are read in the row for $n = j$, and the exponent for x is (n–d, where $d = 2r$. The complete equation for the polynomial is therefore:

stitution of (–x) for x in each of these equations.

The Generation of Applicable Differential Equations

Functions whose coefficients are based on the relations of the Pascal triangle can readily be converted into generating differential equations. In order to show the procedure, assume that the two coefficients for successive terms of the function are specified in terms of the power series:

$$F(x) = \ldots + A(n, r)\, x^r + A(n, s)\, x^s + \ldots \tag{42}$$

where the difference between r and s is a small integer, typically one or two, and the general forms for the A functions are those of binomial, or Pascal, coefficients. Differentiating the equation twice gives the results:

$$\begin{aligned} F'(x) &= \ldots + r\, A(n, r)\, x^{r-1} + s\, A(n, s)\, x^{s-1} + \ldots \\ F''(x) &= \ldots + r(r-1)\, A(n, r)\, x^{r-2} + s(s-1)\, A(n, s)\, x^{s-2} + \ldots \end{aligned} \tag{43}$$

Now, these equations (42) and (43) may be summed in a manner that will reduce the coefficients of each power of x identically to zero. The relation required to do this specifies the differential equation.

The general form for the differential equation, in simplest form for evaluation, is:

$$(a + bx + cx^2)\, y'' + (d + ex)\, y' + (f + gx + hx^2)\, y = 0 \qquad (44)$$

for the conditions under consideration, namely those for which $|r-s| \leqq 2$. Generally, with this equation, either the coefficient of y" is a numeric, or the coefficient of y is a numeric, and the other is then a function of the variable x. (Of course, for the simplest cases, all three coefficients are numerics.)

In the instance that $|r-s| = 1$, then either a, g, and h are zero, or b, d, and h may be zero. In the first of these two cases, if the relation between the coefficients is:

$$A_r / A_{r-1} = -[br(r-1) + dr] / [c(r-1)(r-2) + e(r-1) + f] \qquad (45)$$

then the various coefficients, b, c, d, e, and f, may be isolated and re-inserted in the general differential equation (44). In the second case, the coefficient relation takes the form:

$$A_r / A_{r-1} = [-g] / [cr(r-1) + er + f] \qquad (46)$$

and again the values of c, e, f, and g may be re-inserted in (44).

Frequently, the value of the difference, $|r-s|$ is two, rather than unity. In this instance, the relations corresponding to (45) and (46) take slightly different form. For the first case,

$$A_r / A_{r-1} = -[c(r-2)(r-3) + e(r-2) + f] / [ar(r-1)] \qquad (47)$$

where the values of b, d, g, and h are all zero. Similarly, for the second case, a, b, d, and g are zero, and the relation takes the form:

$$A_r / A_{r-1} = [-h] / [cr(r-1) + er + f] \qquad (48)$$

Generally, the resulting differential equations are hypergeometric differential equations.

The form of the coefficients separated in Eqs. 45 through 48 must agree with the form shown in the corresponding Eq. 42. For example, if in some instance, $r = 2m$, then the terms involving r in Eq. 46 will read $[2cm(2m-1) + 2em + f]$. Similarly, in Eq. 45, the first term in the denominator would read $c(2m-1)(2m-2)$ rather than $c(r-1)(r-2)$. Similarly, if the exponent is $(n-r)$ instead of r, then this term would read $c(n-r-1)(n-r-2)$.

Conclusions

The binomial coefficients tabulated in the Pascal triangle can be useful in many routine calculations encountered in network theory, since their expressions are fundamental to the basic differential equations which apply to network theory.

Bibliography:

Ku, Y. H.:and M. Drubin, Network Synthesis Using Legendre and Hermite Polynomials, J. F. I., Feb. 1962, p. 138

Ahmed, N., and S. Karni, An Application of the Hypergeometric Series in Synthesis Techniques, IEEE International Convention Record, Part 2, 1963

Hamming, R. W., Numerical Methods for Scientists and Engineers, McGraw-Hill Book Co., 1962

Ford, L. R., Differential Equations, McGraw-Hill Book Co., 1933

Wall, H. S., Continued Fractions, Van Nostrand, 1948

Whittaker, E. T., and G. N. Watson, A Course of Modern Analysis, Cambridge University Press, 1927

Jahnke, E., and F. Emde, Funktionentafeln, Dover, 1943

Rideout, V., Active Networks, Prentice-Hall, 1952

Smythe, W. R., Static and Dynamic Electricity, McGraw-Hill Book Co. 1939

Pullen, K. A., Handbook of Transistor Circuit Design, p. 266, Prentice-Hall, Inc. 1961

Pullen, K. A., On the Properties of Ladder Networks, BRL Report 1102, Aberdeen Proving Ground, Md., 1960

Pullen, K. A., On the Relation of the Pascal Triangle to Certain Functions, BRL Report 1214, Aberdeen Proving Ground, 1963

Lance, G. N., Numerical Methods for High-Speed Computers, Iliffe and Sons, Ltd., London, 1960

Masaitis, C., Unpublished communication, 1963

Chrystal, G., Textbook of Algebra, Chelsea Publishing Co., 1952, V. 1, p. 67

Vorob'ev, N. N., Fibonacci Numbers, Blaisdell Publishing Co. NYC. 1961

Jackson, D., Fourier Series and Orthogonal Polynomials, Carus Mathematical Monographs, Mathematical Association of America, 1941

Krall, H. L., and O. Frink, A New Class of Orthogonal Polynomials, the Bessel Polynomials, Trans. Amer. Math. Soc., Vol. 65, 1949, p. 100

Sansome, G., Orthogonal Polynomials, Interscience Publishers, 1959

Hall, H. S., and S. R. Knight, Higher Algebra, McMillan and Co, London, 1945

A HOMOGENEOUS EQUIVALENT OF NONHOMOGENEOUS LINEAR STATE MODELS AND ITS APPLICATION TO THE ANALYSIS OF CONTINUOUS AND DISCRETE-STATE SYSTEMS

H. E. KOENIG and Y. TOKAD
Michigan State University
East Lansing, Michigan

Abstract

This paper presents a procedure for transforming nonhomogeneous systems of differential and difference equations to an equivalent system of homogeneous equations, along with a discussion of some of the basic properties of resulting systems of homogeneous equations and their applications in general system theory.

I Introduction:

General methods for obtaining analytical solutions to linear, constant-coefficients, systems of differential and difference equations as encountered in state-space models of linear time-invarient systems are well established and may be based on either transform calculus (such as the Laplace and z-transforms) or on the properties of functions of a matrix.[1,2] The general forms of the solutions to the continuous and discrete state systems

$$\frac{d}{dt} X(t) = AX(t) + BE(t), \; X(0) = X_o \tag{1}$$

and

$$Y(n+1) = \underline{A}\, Y(n) + \underline{B}\, U(n) \tag{2}$$

as given in terms of functions of a matrix, are respectively

$$X(t) = e^{At} X(\theta) + \int_o^t e^{A(t-\tau)} BE(\tau)\, d\tau \tag{3}$$

and

$$Y(n) = \underline{A}^n Y(0) + \sum_{j=0}^{n-1} \underline{A}^j \underline{B}\; U(n-j-1) \tag{4}$$

where, by definition,

$$e^{At} = U + At + A^2 \frac{t^2}{2!} + \ldots \tag{5}$$

In the special case of homogeneous systems, the convolution integral and convolution summation are, of course, not present and the solutions are given very simply in terms of the eigenvalues and constituent matrices of the coefficient matrices A and $\underline{A}$, i.e., if A has eigenvalues, $\lambda_1, \lambda_2, \ldots \lambda_n$ each of multiplicity k_i, $(i = 1, 2, \ldots n)$, then any function f(A) of the matrix A has the form

$$\begin{aligned} f(A) = {} & Z_{11} f(\lambda_1) + Z_{12} f^1(\lambda_1) + \ldots + Z_{1k_1} f^{(k_1-1)}(\lambda_1) \\ & + Z_{21} f(\lambda_2) + Z_{22} f^1(\lambda_2) + \ldots + Z_{2k_2} f^{(k_2-1)}(\lambda_2) \\ & + \\ & \vdots \\ & + Z_{n1} f(\lambda_n) + Z_{n_2} f^1(\lambda_n) + \ldots + Z_{nk_n} f^{(k_n-1)}(\lambda_n) \end{aligned} \tag{6}$$

where $f^n(\lambda)$ represents the n^{th} derivative of $f(\lambda)$ with respect to λ and $f(\lambda)$ represents the function f(A) with the argument A replaced by λ. The matrices Z_{ij} are a function only of the matrix A, and can be evaluated by any one of several procedures [1], one of which is identical to the partial fraction expansion used in evaluating the inverse Laplace transform or z-transform of a rational function of s or z.

Although, the convolution integral and convolution summation required in the solutions indicated in (3) and (4) can be evaluated by expanding the indicated functions of A and $\underline{A}$ in the form given in (6), the procedure is not always a particularly convenient one to carry out. This is true whether the solution is realized analytically or carried out numerically by machine. The numerical solution to the homogeneous part of (3), for example, can be effectively realized by simply setting t = h, where h is a suitably selected small increment of time, and evaluating e^{Ah} as the first several terms in the infinite series given in (5). The solution is then generated by applying the very simply recursion formula

$$X[h(n+1)] = (e^{Ah})^n X(nh), \qquad n = 1, 2, \ldots \tag{7}$$

The size of the increment h, depends, of course, upon how fast the series in (5) converges and the number of terms one includes in the truncated series. To include the convolution integral obviously involves a much more elaborate computational procedure.

This paper gives a very simple, but very practical procedure for transforming nonhomogeneous systems of differential and difference equations to an equivalent system of homogeneous equations, along with a discussion of some of the basic properties of the resulting systems of homogeneous equations and their applications in general system theory.

II Homogeneous Equivalent Differential Equations

Consider the nth order system of linear homogeneous equations given in (1) and let E(t) be an mth order vector function of the independent variable t, having at most a finite number of first-order discontinuities over the interval I: $(t_1 < t < t_2)$. The infinite series of vectors given in (8) below includes the Fourier series as a special case, and therefore, can be used to represent E(t) everywhere on I except at the finite number of discontinuities, i.e., let

$$E(t) = E_1(t) + E_2(t) + \ldots + E_j(t) + \ldots \tag{8}$$

where

$$E_j(t) = [K_j^o \;\; K_j^{(1)} \;\; \ldots \;\; K_j^{q_j}] \begin{bmatrix} 1 \\ t \\ \vdots \\ t^{q_j} \end{bmatrix} e^{a_j t} \tag{9}$$

$$= K_j F_j(t)$$

and K_j is an n x q_j matrix of constants and $F_j(t)$ is a vector of order q_j. The coefficients in K_j and the exponent α_j in the vector $F_j(t)$ may be real or complex.

Clearly, by setting $\alpha_j = 0$ this series includes the power series in t of order q_j as a special case. Using the series expansion for E(t) given in (8), the system of equations in (1) becomes

$$\frac{d}{dt} X(t) = AX(t) + [BK_1 \quad BK_2 \ \ldots] \begin{bmatrix} F_1(t) \\ F_2(t) \\ \vdots \end{bmatrix} \tag{10}$$

Each vector $F_j(t)$ in (10) represents the solution to a system of homogeneous differential equations

$$\frac{d}{dt} F_j(t) = T_j F_j(t), \quad F_j(0) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{11}$$

where T_j is the bidiagonal matrix

$$T_j = \begin{bmatrix} \alpha_j & q_j & 0 & \ldots & 0 & 0 \\ 0 & \alpha_j & (q_j-1) & \ldots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \ldots & \alpha_j & 1 \\ 0 & 0 & 0 & \ldots & 0 & \alpha_j \end{bmatrix} \tag{12}$$

Augmenting (10) to include (11), the homogeneous equivalent of (1) is

$$\begin{bmatrix} X(t) \\ F_1 \\ F_2 \\ \vdots \\ F_r \\ \vdots \end{bmatrix} = \begin{bmatrix} A & BK_1 & BK_2 & \ldots & BK_r & \ldots \\ 0 & T_1 & 0 & \ldots & 0 & \ldots \\ 0 & 0 & T_2 & \ldots & 0 & \ldots \\ \vdots & & & & & \\ 0 & 0 & 0 & \ldots\ldots & T_r & \ldots \\ \vdots & & & & & \end{bmatrix} \begin{bmatrix} X(t) \\ F_1(t) \\ F_2(t) \\ \vdots \\ F_r(t) \\ \vdots \end{bmatrix}$$

which, for convenience, is written as

$$\frac{d}{dt} X'(t) = A'X'(t) \tag{13}$$

In application, the series given in (8), of course, is truncated after a finite number of terms, depending on the rate of convergence, to give an equivalent homogeneous system of finite order.

The solution to the system of equations in (13) has the form

$$X'(t) = e^{A't} X'(0).$$

Since T_j is bidiagonal, it follows that the characteristic polynomial of A' is given by

$$|\lambda U - A'| = |\lambda U - A| \, (\lambda - a_1)^{q_1} (\lambda - a_2)^{q_2} \ldots (\lambda - a_r)^{q_r}$$

and the order of the characteristic polynomial to be factored in the solution of the homogeneous equations is identical to the order of the characteristic polynomial in the nonhomogeneous system, i.e., the characteristic polynomial has been augmented by a set of known roots. A comparison with the solution to the original system of nonhomogeneous equations by Laplace transform procedures will show that the added factors are in direct correspondence with the zeros introduced by the Laplace transform of the driving functions.

Example 1:

Consider the homogeneous equivalent of (1) when

$$E(t) = \begin{bmatrix} E_1 \cos(\omega t + \theta_1) \\ E_2 \cos(\omega t + \theta_2) \\ \vdots \\ E_m \cos(\omega t + \theta_m) \end{bmatrix} \tag{13}$$

Let (13) be written as

$$E(t) = \frac{1}{2} E_1 e^{j\omega t} + \frac{1}{2} E_1 e^{-j\omega t}$$

where the jth entry in E_1 is the complex number

$$E_j = E_j e^{j\theta_k}$$

Since $e^{\pm j\omega t}$ is realized as a solution to the scalar equation

$$\frac{d}{dt} f(t) = \pm j\omega f(t), \quad f(0) = 1$$

it follows that the augmented system of homogeneous equations is

$$\frac{d}{dt}\begin{bmatrix} X(t) \\ \hline f_1(t) \\ f_2(t) \end{bmatrix} = \left[\begin{array}{c|cc} A & 1/2BE_1 & 1/2BE_1 \\ \hline 0 & j\omega & 0 \\ 0 & 0 & -j\omega \end{array}\right]\begin{bmatrix} X(t) \\ \hline f_1(t) \\ f_2(t) \end{bmatrix}, \quad \begin{bmatrix} X(0) \\ \hline 1 \\ 1 \end{bmatrix} \tag{14}$$

or

$$\frac{d}{dt} X'(t) = A' X(t)$$

If the eigenvalues of A, for example, are distinct and not equal to $\pm j\omega$, i.e., if the eigenvalues of A' are distinct, then the solution to (14) as obtained by expanding A' into its constituent matrix is of the form

$$\begin{bmatrix} X(t) \\ f_1(t) \\ f_2(t) \end{bmatrix} = e^{A't}\begin{bmatrix} X(0) \\ 1 \\ 1 \end{bmatrix}$$

or

$$\begin{bmatrix} X(t) \\ \hline f_1(t) \\ f_2(t) \end{bmatrix} = \left[\begin{array}{c|cc} Z^{(1)}_{11} & Z^{(1)}_{12} & Z^{(1)}_{13} \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right]\begin{bmatrix} X(0) \\ \hline 1 \\ 1 \end{bmatrix} e^{\lambda_1 t} + \ldots + \left[\begin{array}{c|cc} Z^{(n)}_{11} & Z^{(n)}_{12} & Z^{(n)}_{13} \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right]\begin{bmatrix} X(0) \\ \hline 1 \\ 1 \end{bmatrix} e^{\lambda_n t}$$

$$+\left[\begin{array}{c|cc} 0 & Z_{12}^{(n+1)} & Z_{13}^{(n+2)} \\ \hline 0 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right]\left[\begin{array}{c} X(0) \\ \hline 1 \\ 1 \end{array}\right] e^{j\omega t} + \left[\begin{array}{c|cc} 0 & Z_{12}^{(n+2)} & Z_{13}^{(n+2)} \\ \hline 0 & 0 & 0 \\ 0 & 0 & 1 \end{array}\right]\left[\begin{array}{c} X(0) \\ \hline 1 \\ 1 \end{array}\right] e^{-j\omega t} \tag{15}$$

where $Z_{12}^{(j)} = \hat{Z}_{13}^{(j)}$, $j = 1, a, \ldots n$, $Z_{12}^{(n+1)} = \hat{Z}_{12}^{(n+2)}$, $Z_{13}^{(n+1)} = \hat{Z}_{13}^{(n+2)}$ and all matrices are idempotent and sum to unity. That the solution is of the form indicated follows directly from the general form of the expression given in (6). For example, to investigate the general properties of the last constituent matrix in (15), take

$$f(\lambda) = (\lambda - \lambda_1)\ (\lambda - \lambda_2)\ \ldots\ (\lambda - \lambda_1)\ (\lambda - j\omega)$$

The solution of interest, of course, if obtained by partitioning the constituent matrices as indicated.

If the constituent matrices in (15) are evaluated by means of partial fraction expansion, it isfound that exactly the same steps are required in evaluating $e^{A't}$ as are required in evaluating the inverse Laplace transform of

$$X(s) = (sU-A)^{-1}\,BE(s) \tag{16}$$

by partial fraction expansion. To evaluate (16) by means of the convolution integral is equivalent to solving (1) as a system of non-homogeneous equations.

It should be evident that when the excitation vector is discontinuous, an equivalent homogeneous system can be developed for each continuous region. In the case of periodic functions, an equivalent homogeneous system is developed for each continuous subregion in one period. In general, for such simple functions as pulses, triangular waves, etc., there are only very minor changes in the augmented matrix A' from one continuous interval to the next. The corresponding changes in the constituent matrices are also usually only minor.

III Homogeneous Equivalents of Difference Equations

Homogeneous equivalents of nonhomogeneous systems of difference equations are exactly similar to those of the differential equations. The general form of the polynomial used to approximate the sequence of inputs in (2) is taken as

$$U(n) = U(n) + U_2(n) + \dots$$

where

$$U_j(n) = [K_j \quad K_j^{(1)} \quad \dots \quad K_j^{(q_j)}] \begin{bmatrix} 1 \\ n \\ \vdots \\ n^{q_j} \end{bmatrix} \alpha^n \tag{17}$$

$$= K_j \, F_j(n) \tag{18}$$

When $\alpha = 1$ the series represents a power series in n.

Each term in the series of (17) represents the solution to the difference equations

$$F_j(n+1) = T_j \, F_j(n), \quad F_j(0) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

where T_j is the triangular matrix

$$T_j = \alpha_j \begin{bmatrix} 1 & q_j & \dots & q_j & 1 \\ 0 & 1 & \dots & q_j - 1 & 1 \\ \vdots & & & & \\ 0 & 0 & \dots & 2 & 1 \\ 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

In general, the non zero entries in row i of T_j are the coefficients of the polynomial $(q_j + 1)^i$.

Example:

Consider the homogeneous equivalent of (2) when

$$U(n) = U_o + U_1 n + U_2 n^2$$

The required power series is generated as a solution to the third order system of equations and the equivalent homogeneous system is

$$\begin{bmatrix} X(n+1) \\ \hline f_2(n+1) \\ f_1(n+1) \\ f_o(n+1) \end{bmatrix} = \left[\begin{array}{c|ccc} A & BU_2 & BU_1 & BU_o \\ \hline 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{array}\right] \begin{bmatrix} X(n) \\ \hline f_2(n) \\ f_1(n) \\ f_o(n) \end{bmatrix} , \quad \begin{bmatrix} X(0) \\ \hline f_2(0) \\ f_1(0) \\ f_o(0) \end{bmatrix} = \begin{bmatrix} X_o \\ \hline 0 \\ 0 \\ 1 \end{bmatrix}$$

References:

1. Gantmacher, F.R., "The Theory of Matrices" Vol. 1, Chelsea Publishing Co., 1959.
2. Zadeh, L.A., and Desoer, C.A., "Linear System Theory, The State Approach", McGraw-Hill, 1963.

OPTIMUM DESIGN OF LINEAR PULSE-WIDTH-MODULATED CONTROL SYSTEMS

T. MEKSAWAN and G. J. MURPHY
Information-Processing and Control
Systems Laboratory
Northwestern University
Evanston, Illinois

Abstract

A technique for the synthesis of linear control systems with pulse-width modulation which are optimal according to a specified quadratic performance criterion is described in this paper. The technique leads to the design of a linear digital controller. The resulting system exhibits zero error at sampling instants and approximately dead-beat performance after the first few sampling periods. The synthesis procedure is developed by the use of the state-transition method, and its use is illustrated by means of a specific example.

Introduction

In a previous paper[7], a technique for the synthesis of linear pulse-width-modulated computer-controlled systems to fulfill an optimal performance criterion was presented by the authors. The design technique is directed toward the determination of the optimum sequence of input signals to the pulse-width modulator. The optimum value of the modulator input is calculated at each sampling instant from the state variables of the system by use of a digital computer incorporated within the control system. For a step input, the optimum second-order system is characterized by zero error at all sampling instants except the one at which the step function is applied and by approximately dead-beat response after the first few sampling periods. The system response between sampling instants can be modified by adjustment of either the pulse amplitude or the gain constant of the system.

For a practical control system, a simple method of control is desirable. The use of a digital computer within the control loop of the system is usually very expensive and hence is to be avoided, if possible. The purpose of this

---

The work described in this paper was supported by the Air Force Office of Scientific Research.

paper, therefore, is to present a useful and straightforward procedure for the synthesis of a discrete compensating network for such a system. In this design method the need for a high-speed digital computer within the control loop of the system is eliminated. Such a computer can, however, be used to advantage in the course of synthesizing the discrete compensator or digital controller for optimal control.

Mathematical Description of the System

A block diagram of the system under consideration is shown in Fig. 1. The plant or process to be controlled is assumed to be described by an n th - order ordinary linear differential equation with constant coefficients, and the output of the pulse-width modulator is

$$m(t) = \begin{cases} m(jT), & 0 \leq (t - jT) < \tau_j \\ 0, & \tau_j \leq (t - jT) < T \end{cases} \tag{1}$$

$$j = 0, 1, 2, \ldots,$$

where τ_j is the pulse width during the j th sampling interval, and T is the sampling period. The modulator output is related to the modulator input $e_1(t)$ by

$$m(jT) = M \operatorname{sgn} e_1(jT), \quad j = 0, 1, 2, \ldots \tag{2}$$

and

$$\tau_j = T \operatorname{sat} \left[\left(\frac{\alpha}{T} \right) \left| e_1(jT) \right| \right], \quad j = 0, 1, 2, \ldots, \tag{3}$$

in which α is a positive constant of proportionality, M is a positive constant, and the signum and saturation functions are defined by

$$\operatorname{sgn}(x) = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \\ 0, & x = 0 \end{cases} \tag{4}$$

and

$$\text{sat}(x) = \begin{cases} 1 , & x \geq 1 \\ x , & 0 \leq |x| < 1 \\ -1 , & x \leq -1 \end{cases} . \tag{5}$$

The state vectors are denoted by

$$\underset{\sim}{x} = [x_1 , x_2 , \ldots , x_r]' , \tag{6}$$

$$\underset{\sim}{y} = [y_1 , y_2 , \ldots , y_n]' , \tag{7}$$

and

$$\underset{\sim}{w} = [x' , y']' , \tag{8}$$

where the prime denotes the transpose, where x_1 and y_1 denote the input and the output, respectively, of the system, and where x_j and y_k are the other components of the input vector and the output vector, respectively, for $j = 2, 3, \ldots , r$ and $k = 2, 3, \ldots , n$.

The system is characterized by the differential equations

$$\dot{\underset{\sim}{w}} (t) = \underset{\sim}{A} \underset{\sim}{w} (t) + \underset{\sim}{b} m (jT) , \quad 0 \leq (t - jT) < \tau_j , \quad j = 0, 1, 2, \ldots \tag{9}$$

and

$$\dot{\underset{\sim}{w}} (t) = \underset{\sim}{A} \underset{\sim}{w} (t) , \quad \tau_j \leq (t - jT) < T , \quad j = 0, 1, 2, \ldots . \tag{10}$$

By combining the solutions to these two equations, one obtains

$$\underset{\sim}{w} [(j + 1) T] = \underset{\sim}{\Phi} (T) \underset{\sim}{w} (jT) + \underset{\sim}{d} (\tau_j) m (jT) , \quad j = 0, 1, 2, \ldots , \tag{11}$$

in which

$$\underset{\sim}{\Phi} (T) = \exp (T \underset{\sim}{A}) \tag{12}$$

and

$$\underset{\sim}{d} (\tau_j) = \underset{\sim}{\Phi} (T) \int_0^{\tau_j} \underset{\sim}{\Phi} (- \tau) \underset{\sim}{b} \, d \tau . \tag{13}$$

A quadratic loss criterion of the form

$$I = \sum_{j=1}^{\infty} [x_1 (jT) - y_1 (jT)]^2 \tag{14}$$

is chosen as a measure of the performance of the system. Evidently the minimum value

$$I_{min} = 0 \tag{15}$$

of I will be realized if and only if

$$y_1 [(j + 1) T] = x_1 [(j + 1) T] , \quad j = 0, 1, 2, \ldots . \tag{16}$$

From (11), it is found that

$$y_1[(j + 1)T] = \sum_{k=r+1}^{r+n} \phi_{r+1,k}\, y_{k-r}(jT) + m\,(jT)\, d_{r+1}\,(\tau_j) , \quad j = 0, 1, 2, \ldots . \tag{17}$$

If $\underset{\sim}{x}(t)$ is deterministic, $x_1[(j + 1) T]$ can be specified at $t = jT$, $j = 0, 1, 2, \ldots$. Therefore, the minimum value of I will be realized by making

$$m(jT)\, d_{r+1}\,(\tau_j) = y_1[(j + 1)T] - \sum_{k=r+1}^{r+n} \phi_{r+1,k}\, y_{k-r}\,(jT) , \quad j = 0, 1, 2, \ldots , \tag{18}$$

if this can be done.

By virtue of (2) ,

$$\left| m\,(jT)\right| = M . \tag{19}$$

It follows that the minimum value of I will be realized if M and τ_j , $j = 0, 1, 2, \ldots$, are so chosen that

$$M \left| d_{r+1}(\tau_j)\right| = \left| y_1[(j + 1)T] - \sum_{k=r+1}^{r+n} \phi_{r+1,k}\, y_{k-r}(jT)\right| , \quad j = 0, 1, 2, \ldots , \tag{20}$$

subject to the constraint

$$0 \leq \tau_j \leq T , \quad j = 0, 1, 2, \ldots , \tag{21}$$

and if then sgn m(jT) is chosen in accordance with (18).

The existence of solutions for τ_j, $j = 0, 1, 2, \ldots$, that satisfy (21) depends on the value of M; the manner in which the value of M can be selected to ensure the existence of such solutions (if there are no constraints on M) is indicated later in this paper in an illustrative example. If there are constraints on M (due to saturation, for example), such solutions may or may not exist; whether they do indeed exist, and, if so, for what range of allowable values of M, can be determined by the method illustrated also in the illustrative example.

If such solutions for τ_j, $j = 0, 1, 2, \ldots$, exist, let them be denoted by τ_j^{o}, $j = 0, 1, 2, \ldots$, and then also let $\underset{\sim}{y}^{o}(jT)$ denote the state $\underset{\sim}{y}(jT)$ that is obtained if

$$\tau_i = \tau_i^{o}, \quad i = 0, 1, 2, \ldots, j-1 . \qquad (22)$$

Development of the Method

The application of digital controllers or sampled-data processing units in discrete-data control systems to fulfill an optimal performance criterion has been a subject of considerable study in recent years[1,2,3,4,5,6]. A similar technique can be employed successfully in the design of optimal linear systems with pulse-width modulation. The synthesis procedure here must be directed toward the determination of the optimum input- and output-sequences of the pulsed-data network. A pulse transfer function can then be found, and thus the required digital controller is defined.

The optimum sequence of input signals to the pulse-width modulator is given by

$$e_1^{o}(jT) = \frac{1}{\alpha}\tau_j^{o} \operatorname{sgn} m(jT), \quad j = 0, 1, 2, \ldots, \qquad (23)$$

where α is defined as in (3), for the case under discussion. The Z-transform of this sequence (regarded as the output of an ideal sampler) is

$$E_1^{o}(z) = e_1^{o}(0) + e_1^{o}(T)\, z^{-1} + e_1^{o}(2T)\, z^{-2} + \ldots \quad . \qquad 24)$$

When $e_1^{o}(jT)$ is known for $j = 0, 1, 2, \ldots$, the optimal response of the system can be determined. That is, the sequence of optimal values of the system state variables at successive sampling instants can be calculated by the use of the transition equation

$$\underset{\sim}{w}^o\,[(j+1)T] = \underset{\sim}{\Phi}(T)\,\underset{\sim}{w}^o(jT) + \underset{\sim}{d}(\tau_j^{\,o})\,m(jT)\,, \qquad j = 0, 1, 2, \ldots \quad . \tag{25}$$

The sequence of optimal input signals to the pulsed-data network can then be expressed as

$$e_2^{\,o}(jT) = \underset{\sim}{\beta}'\,\underset{\sim}{w}^o(jT)\,, \qquad j = 0, 1, 2, \ldots\,, \tag{26}$$

where $\underset{\sim}{\beta}$ is a constant vector. The choice of the components of $\underset{\sim}{\beta}$ can be made arbitrarily simply, which is very desirable in dealing with systems with output constraints, where not all components of the system state vector are available physically for observation. Finally, the optimal transfer function of the required digital controller is found to be

$$D^o(z) = \frac{E_1^{\,o}(z)}{E_2^{\,o}(z)} = \frac{E_1^{\,o}(z)}{Z\,[\underset{\sim}{\beta}'\,\underset{\sim}{w}^o(jT)]}\,, \tag{27}$$

which is physically realizable unless $e_2^{\,o}(0) = 0$. A discussion of techniques for the physical realization of $D^o(z)$ can be found elsewhere [3,9,10].

This synthesis procedure can also be applied to the design of digital controllers for linear and nonlinear discrete-data control systems in general. It is shown in another paper[8] by the authors that the synthesis of pulse transfer functions by this method yields the same results as various techniques described elsewhere in the literature[6,10].

Optimal Control of Second-Order Plants

Consider the case of an initially quiescent second-order linear process with the transfer function

$$G(s) = \frac{Y_1(s)}{M(s)} = \frac{K}{s(s+a)} \; . \qquad (28)$$

where K > 0 and a > 0 are fixed and known. Let the input to the pulse-width modulated control system be denoted by $x_1(t)$, and

$$\underset{\sim}{w}(t) = [\, x_1(t) \, , \, y_1(t) \, , \, y_2(t) \,]' \, , \qquad (29)$$

where

$$y_2(t) \triangleq \dot{y}(t) \, . \qquad (30)$$

The state diagram of this process is shown in Fig. 2, for the case in which $x_1(t)$ is a step function. For $T = a = K = x_1(0) = 1$, it is found[7] that the optimum response of this system, with respect to the criterion given in (14), is described by Table I* .

TABLE I .

j	m(jT)	τ_j^o	$\alpha e_1^o(jT)$	$x_1(jT)$	$y_1^o(jT)$	$y_2^o(jT)$
0	2.729	0.946	0.946	1	0	0
1	-2.729	0.946	-0.946	1	1	1.582
2	2.729	0.428	0.428	1	1	-1.000
3	-2.729	0.063	-0.063	1	1	0.169
4	2.729	0.001	0.001	1	1	-0.003
5	-	0.000	0.000	1	1	0.000

If, now, one chooses the vector $\underset{\sim}{\beta}$ in (26) so that

$$\underset{\sim}{\beta}' = [\, 1 \quad -1 \quad \tfrac{1}{2} \,] \, , \qquad (31)$$

then one obtains five non-vanishing values of $e_2^o(jT)$:

$$e_2^o(0) = 1, \quad e_2^o(T) = 0.791, \quad e_2^o(2T) = -0.500 \, , \quad e_2^o(3T) = 0.0850 \, ,$$
$$e_2^o(4T) = -0.002 \, . \qquad (32)$$

---

*The numerical results presented in this table are approximations obtained by solving transcendental equations.

It follows that the required pulse transfer function is

$$D^{o}(z) = \left(\frac{1}{\alpha}\right) \frac{(0.946 - 0.946z^{-1} + 0.428z^{-2} - 0.063z^{-3} + 0.001z^{-4})}{(1 + 0.791z^{-1} - 0.500z^{-2} + 0.0850z^{-3} - 0.002z^{-4})} \cdot \tag{33}$$

A block diagram of the complete system thus obtained is shown in Fig. 3 .

Control of a Third-Order Process

Consider next the case of an initially quiescent third-order linear process with the transfer function

$$G(s) = \frac{K}{s\,(s + a_1)(s + a_2)} \quad , \tag{34}$$

where $K > 0$ and $a_1 > a_2 > 0$ are fixed and known. A state diagram of this process is shown in Fig. 4, for the case in which the input, $x_1(t)$, is a step function.

With

$$\underset{\sim}{w} \overset{\Delta}{=} [\, x_1, \ y_1, \ y_2, \ y_3 \,]' \quad , \tag{35}$$

inspection of the state diagram yields

$$\underset{\sim}{\Phi}(T) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \phi_{23}(T) & \phi_{24}(T) \\ 0 & 0 & \phi_{33}(T) & \phi_{34}(T) \\ 0 & 0 & 0 & \phi_{44}(T) \end{bmatrix} \tag{36}$$

and

$$d(\tau_j) = \begin{bmatrix} 0 \\ d_2(\tau_j) \\ d_3(\tau_j) \\ d_4(\tau_j) \end{bmatrix} \quad , \tag{37}$$

where

$$\phi_{23}(T) = \left(\frac{1}{a_2}\right)(1 - e^{-a_2 T}) \ , \tag{38}$$

$$\phi_{24}(T) = \frac{1}{(a_1 - a_2)} \left[\left(\frac{1}{a_2} - \frac{1}{a_1}\right) + \frac{1}{a_1} e^{-a_1 T} - \frac{1}{a_2} e^{-a_2 T} \right] \ , \tag{39}$$

$$\phi_{33}(T) = e^{-a_2 T} \quad , \tag{40}$$

$$\phi_{34}(T) = \frac{1}{(a_1 - a_2)} \left(e^{-a_2 T} - e^{-a_1 T} \right) , \tag{41}$$

$$\phi_{44}(T) = e^{-a_1 T} , \tag{42}$$

$$d_2(\tau_j) = \frac{K}{(a_1 - a_2)} \cdot \Big[\left(\frac{1}{a_2} - \frac{1}{a_1} \right) \tau_j + \left(\frac{1}{a_2^{\,2}} \right)\left(e^{-a_2 T} \right) - \left(\frac{1}{a_1^{\,2}} \right)\left(e^{-a_1 T} \right) + \left(\frac{1}{a_1^{\,2}} \right)\left(e^{-a_1 T} \right) e^{a_1 \tau_j} - \left(\frac{1}{a_2^{\,2}} \right)\left(e^{-a_2 T} \right) e^{a_2 \tau_j} \Big] , \tag{43}$$

$$d_3(\tau_j) = \frac{K}{(a_1 - a_2)} \Big[\left(\frac{1}{a_1} \right)\left(e^{-a_1 T} \right) - \left(\frac{1}{a_2} \right)\left(e^{-a_2 T} \right) - \left(\frac{1}{a_1} \right)\left(e^{-a_1 T} \right) e^{a_1 \tau_j} + \left(\frac{1}{a_2} \right)\left(e^{-a_2 T} \right) e^{a_2 \tau_j} \Big] , \tag{44}$$

and

$$d_4(\tau_j) = \left(\frac{K}{a_1} \right)\left(e^{-a_1 T} \right)\left(e^{a_1 \tau_j} - 1 \right) . \tag{45}$$

Substituting (36) - (45) into (18) gives

$$m(jT)\, d_2(\tau_j^{\,o}) = x_1(jT) - y_1^{\,o}(jT) - \phi_{23}(T)\, y_2^{\,o}(jT) - \phi_{24}(T)\, y_3^{\,o}(jT), \quad j = 0, 1, 2, \dots , \tag{46}$$

or

$$\left(\alpha_3\, e^{a_2 \tau_0^{\,o}} - \alpha_2\, e^{a_1 \tau_0^{\,o}} \right) = \tau_0^{\,o} + \alpha_1 - \frac{a_1 a_2\, x_1(0)}{K\, m(0)} \tag{47}$$

and

$$\left(\alpha_3\, e^{a_2 \tau_j^{\,o}} - \alpha_2\, e^{a_1 \tau_j^{\,o}} \right) = \tau_j^{\,o} + \alpha_1 + \frac{a_1\, a_2}{K\, m(jT)} \Big[\phi_{23}(T)\, y_2^{\,o}(jT) + \phi_{24}(T)\, y_3^{\,o}\,(jT) \Big], \quad j = 1, 2, \dots , \tag{48}$$

where

$$\alpha_1 \triangleq \frac{1}{(a_1 - a_2)} \Big[\left(\frac{a_1}{a_2} \right)\left(e^{-a_2 T} \right) - \left(\frac{a_2}{a_1} \right)\left(e^{-a_1 T} \right) \Big] , \tag{49}$$

$$\alpha_2 \triangleq \frac{1}{(a_1 - a_2)} \Big[\left(\frac{a_2}{a_1} \right)\left(e^{-a_1 T} \right) \Big] , \tag{50}$$

and

$$\alpha_3 \triangleq \frac{1}{(a_1 - a_2)} \Big[\left(\frac{a_1}{a_2} \right)\left(e^{-a_2 T} \right) \Big] . \tag{51}$$

The solutions to these equations can be found by finding the intersection of the two curves representing respectively the right- and left-hand members of (47) or (48), as shown in Fig. 5. Inspection of these curves reveals[7] that it is necessary that $m(0)$ have the same sign as $x_1(0)$, that $m(jT)$ and $\phi_{23}(T)\, y_2^{o}(jT) + \phi_{24}(T)\, y_3^{o}(jT)$ be of opposite signs for $j = 1,2, \ldots,$ and that

$$\left| y_1^{o}(T) \right| \geq \left| \phi_{23}(T)\, y_2^{o}(T) + \phi_{24}(T)\, y_3^{o}(T) \right| \tag{52}$$

or

$$\left| d_2(\tau_0^{o}) \right| \geq \left| \phi_{23}(T)\, d_3(\tau_0^{o}) + \phi_{24}(T)\, d_4(\tau_0^{o}) \right| , \tag{53}$$

if the solutions to (47) and (48) are to satisfy the relations

$$T \geq \tau_0 \geq \tau_1 \geq \tau_2 \geq \ldots \tag{54}$$

Substituting the known value of $d(\tau_0^{o})$ and simplifying the result yield the requirement

$$\begin{aligned} \tau_0^{o} \geq \frac{1}{(a_1 - a_2)} \Bigg[& \left(\frac{a_2}{a_1}\right)(e^{-a_1 T})(2-e^{-a_1 T}) - \left(\frac{a_1}{a_2}\right)(e^{-a_2 T})(2-e^{-a_2 T}) \\ & -\left(\frac{a_2}{a_1}\right)(e^{-a_1 T})(2-e^{-a_1 T})\, e^{a_1 \tau_0^{o}} + \left(\frac{a_1}{a_2}\right)(e^{-a_2 T})(2-e^{-a_2 T})\, e^{a_2 \tau_0^{o}} \Bigg] . \end{aligned} \tag{55}$$

This expression can be used to determine the largest permissible value, τ_{0m}^{o}, of τ_0^{o} in the allowable range, $0 \leq \tau_0^{o} \leq T$, as indicated in Fig. 6. The value of τ_0^{o} to be used is taken to be

$$\tau_0^{o} = T \operatorname{sat} \left(\frac{\tau_{0m}^{o}}{T} \right) . \tag{56}$$

The knowledge of this initial optimum pulse width can be used to find the minimum allowable value of the pulse amplitude M, in a manner to be illustrated later.

Synthesis Procedure

The optimum design for the third-order pulse-width-modulated control system described in the preceding section can be carried out in the following steps:

1. Find $\tau_{0\,m}^{o}$ from

$$\tau_{0\,m}^{o} = \frac{1}{a_1 - a_2} \left[e^{-a_1 T} (2-e^{-a_1 T}) \frac{a_2}{a_1} (1-e^{a_1 \tau_{0\,m}^{o}}) - e^{-a_2 T} (2-e^{-a_2 T}) \frac{a_1}{a_2} (1-e^{a_2 \tau_{0\,m}^{o}}) \right] , \qquad (57)$$

and determine the value of τ_0^{o} by use of (56) .

2. Determine the corresponding (minimum) value of $M > 0$ from

$$M = \frac{a_1 a_2 \; | x_1(0) |}{K (\tau_0^{o} + \alpha_1 + \alpha_2 e^{a_1 \tau_0^{o}} - \alpha_3 e^{a_2 \tau_0^{o}})} . \qquad (58)$$

3. Calculate the values of $m(jT)$ and τ_j^{o}, $j = 1, 2, 3, \ldots$, from

$$m(jT) = - M \operatorname{sgn} [\phi_{23}(T)\, y_2^{o}(jT) + \phi_{24}(T)\, y_3^{o}(jT)] \qquad (59)$$

and

$$(\alpha_3 e^{a_2 \tau_j^{o}} - \alpha_2 e^{a_1 \tau_j^{o}}) = \tau_j^{o} + \alpha_1 + \frac{a_1 a_2}{K m(jT)} [\phi_{23}(T)\, y_2^{o}(jT) + \phi_{24}(T)\, y_3^{o}(jT)] ,$$

$$j = 1, 2, \ldots \; . \qquad (60)$$

4. Determine the optimum state of the system at each sampling instant until $y_2^{o}(jT)$ and $y_3^{o}(jT)$ become negligible from

$$\underset{\sim}{w}^{o}[(j + 1)T] = \underset{\sim}{\Phi} (T)\, \underset{\sim}{w}^{o}(jT) + \underset{\sim}{d} (\tau_j^{o})\, m(jT) . \qquad (61)$$

5. Choose the instrumentation vector $\underset{\sim}{\beta}$, and determine $e_2(jT)$, $j = 0, 1, 2, \ldots$, from

$$e_2^{o}(jT) = \underset{\sim}{\beta}' \underset{\sim}{w}^{o}(jT) . \qquad (62)$$

6. Determine $e_1^{o}(jT)$, $j = 0, 1, 2, \ldots$, from

$$e_1^{o}(jT) = \frac{1}{\alpha} \tau_j^{o} \operatorname{sgn} m(jT) . \qquad (63)$$

7. Determine the optimal pulse transfer function for the digital controller from

$$D^{o}(z) = \frac{E_1^{o}(z)}{E_2^{o}(z)} . \qquad (64)$$

Numerical Example

The following numerical example is presented to illustrate the use of the procedure outlined in the preceding section.

Consider a specific process with $T = K/4 = a_1/2 = a_2 = 1$. From (57) it is found that

$$\tau_{0\,m}^{o} = (1.201)\, e^{\tau_{0\,m}^{o}} - (0.126)\, e^{2\tau_{0\,m}^{o}} - 1.075 = 0.14 . \qquad (65)$$

Use of (56) and (58) then yields a value of

$$M = 9.6975 \left| x_1(0) \right| . \qquad (66)$$

For ease in computation, $M = 10 \left| x_1(0) \right|$ will be used instead. The numerical values obtained for a unit step input to the initially quiescent system are given in Table II* . Since the state variables $y_2^{o}(9T)$ and $y_3^{o}(9T)$ are zero, τ_9^{o} is zero, and the desired state has then been reached at time 9T .

The input signal to the pulsed-data network can now be chosen, as in (62). Thus, we write

$$e_2^{o}(jT) = \beta_1 x_1(jT) + \beta_2 y_1^{o}(jT) + \beta_3 y_2^{o}(jT) + \beta_4 y_3^{o}(jT) . \qquad (67)$$

The choice of the instrumentation factors β_1, β_2, β_3, and β_4 now is in-

---

*Again it should be noted that the numerical values are approximations obtained by solving transcendental equations.

TABLE II

j	$m(jT)$	τ_j^o	$\alpha e_1^o(jT)$	$x_1(jT)$	$y_1^o(jT)$	$y_2^o(jT)$	$y_3^o(jT)$
0	10	0.136	0.136	1	0	0	0
1	-10	0.135	-0.135	1	1	1.299	0.847
2	10	0.069	0.069	1	1	-0.610	-0.722
3	-10	0.029	-0.029	1	1	0.259	0.303
4	10	0.011	0.011	1	1	-0.101	-0.118
5	-10	0.004	-0.004	1	1	0.038	0.044
6	10	0.001	0.001	1	1	-0.014	-0.016
7	-10	0.0005	-0.0005	1	1	0.005	0.006
8	10	0.0000	0.0000	1	1	-0.002	-0.002
9	-10	0.0000	0.0000	1	1	0.000	0.000

fluenced by the accessibility of the components of the state vectors. Let it be assumed that $y_3(t)$ does not represent any physical quantity and hence is not measurable. One possible choice will then be $\beta_1 = -\beta_2 = 2\beta_3 = 1$ and $\beta_4 = 0$. The required pulse transfer function for the digital controller is then

$$D^o(z)=\left(\frac{1}{\alpha}\right) \frac{(0.136-0.135z^{-1}+0.069z^{-2}-0.029z^{-3}+0.011z^{-4}-0.004z^{-5}+0.001z^{-6}-0.0005z^{-7})}{(1+0.650z^{-1}-0.305z^{-2}+0.130z^{-3}-0.051z^{-4}+0.019z^{-5}-0.007z^{-6}+0.003z^{-7}-0.001z^{-8})} . \tag{68}$$

A block diagram of this optimum system is shown in Fig. 7. (It should be noted that $D^o(z)$ is not unique, since it depends upon the choice of M and the feedback vector $\underset{\sim}{\beta}$.) Complete system responses for various values of M are shown in Fig. 8.

It is evident that no solution of the kind proposed* exists if (because of saturation, for example) M is limited to the range

---

*That is, a solution for which $I = 0$.

$$0 < M \leq M_{max} < 9.6975 \; |x_1(0)| . \tag{69}$$

If

$$M_{max} \geq 9.6975 \; |x_1(0)| \; , \tag{70}$$

the optimum solution is obtained by choosing

$$M = M_{max} \tag{71}$$

and then determining the corresponding values of $\tau_j^{\,o}$. If

$$M_{max} < 9.6975 \; |x_1(0)| \; , \tag{72}$$

again one first chooses

$$M = M_{max} \; . \tag{71}$$

Then, however,

$$[\, x_1(T) - y_1(T) \,]^2 \neq 0 \tag{73}$$

and hence

$$\min (I) \neq 0 \; . \tag{74}$$

Using (71), one can then compute $\underset{\sim}{w}(T)$, which may be regarded as an initial condition on a new time scale

$$t' \overset{\Delta}{=} t - T \; . \tag{75}$$

The problem is now similar to the original problem; the differences are that a) the system is not initially quiescent, and b) the value of M is fixed at M_{max} .

The procedure is repeated as often as necessary to obtain convergence of the $\tau_j^{\,o}$. The solution will converge for polynomial inputs, as can be seen, if the linear process is stable and of sufficiently high Type number.

Conclusion

By use of the technique described in this paper, it is possible to synthesize a linear digital controller for optimal control of a linear process driven by a pulse-width modulator in a feedback loop. The procedure described is systematic and fairly simple, and it can be programmed on a digital computer, if desired.

The same technique can be used to synthesize a digital controller for the time-optimal control of a linear sampled-data system or of a large class of nonlinear sampled-data systems[8].

References

1. A. R. Bergen and J. R. Ragazzini, "Sampled-Data Processing Techniques for Feedback Control Systems", Trans.AIEE, vol. 73. part II, 1954, pp. 236-47.
2. J. E. Bertram, "Factors in the Design of Digital Controllers for Sampled-Data Feedback Systems", Trans. AIEE, vol. 75, part II, 1956, pp.151-158.
3. J. Sklansky, "Pulsed R C Networks for Sampled-Data Systems", IRE Convention Record, part 2, 1956, pp. 81-96.
4. E. I. Jury and W. Schroeder, "Discrete Compensation for Sampled-Data and Continuous Control Systems", Trans. AIEE, vol. 75, part II, 1956, pp. 317-325.
5. K. K. Maitra and P. E. Sarachik, "Digital Compensation of Continuous-Data Feedback Control Systems", Trans. AIEE, vol. 75, part II, 1956, pp. 107-115.
6. J. T. Tou and B. Vadhanaphuti, "Optimum Control of Nonlinear Discrete-Data Systems", Trans. AIEE, vol. 80, part II, 1961, pp. 166-170.
7. G. J. Murphy and T. Meksawan, "Optimal Synthesis of Linear Pulse-Width-Modulated Computer-Controlled Systems by Dynamic Programming", to be published.
8. T. Meksawan and G. J. Murphy, "Optimum Nonlinear Discrete-Data Control Systems", Regelungstechnik, vol. 7, 1963, pp. 295-299.
9. J. T. Tou, Digital and Sampled-Data Control Systems, McGraw-Hill Book Co., Inc., New York, 1959.
10. E. I. Jury, Sampled-Data Control Systems, John Wiley & Sons, Inc., New York, 1958.

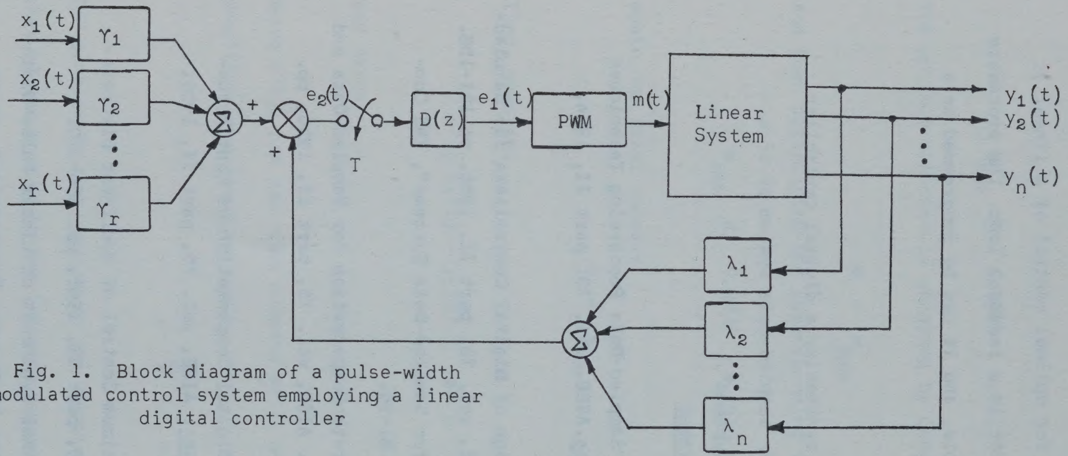

x1(t)
γ1
x2(t)
γ2
xr(t)
γr
Σ
+
+
e2(t)
T
D(z)
e1(t)
PWM
m(t)
Linear System
y1(t)
y2(t)
yn(t)
λ1
λ2
λn
Σ

Fig. 1. Block diagram of a pulse-width modulated control system employing a linear digital controller

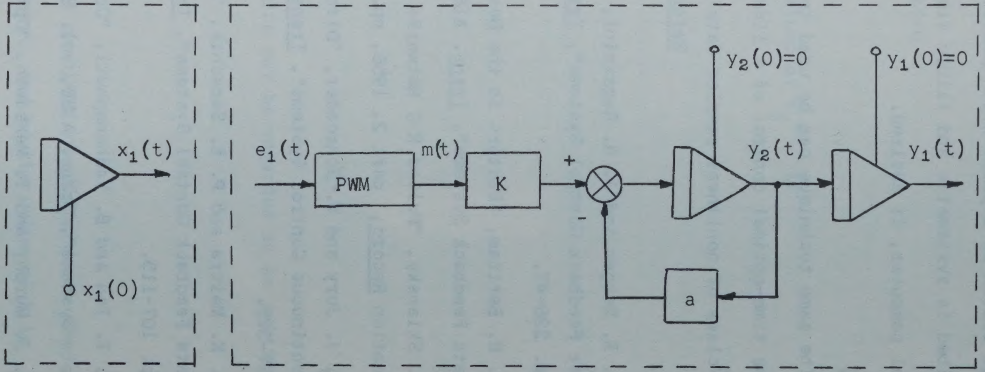

x1(t)
x1(0)
e1(t)
PWM
m(t)
K
+
-
y2(0)=0
y1(0)=0
y2(t)
y1(t)
a

Fig. 2. Representation of a second-order process

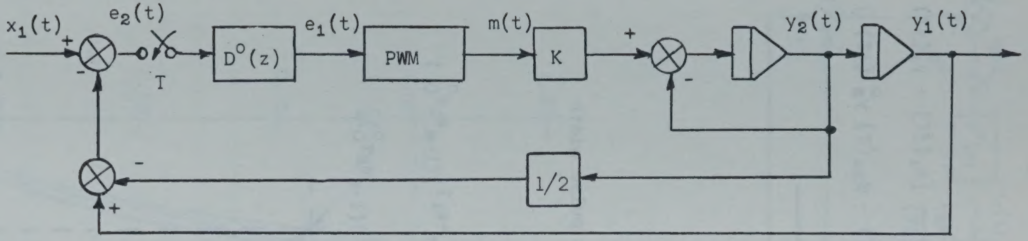

$x_1(t)$
$e_2(t)$
T
$D^o(z)$
$e_1(t)$
PWM
$m(t)$
K
$y_2(t)$
$y_1(t)$
1/2
+
-

Fig. 3. Block diagram of an optimal second-order system

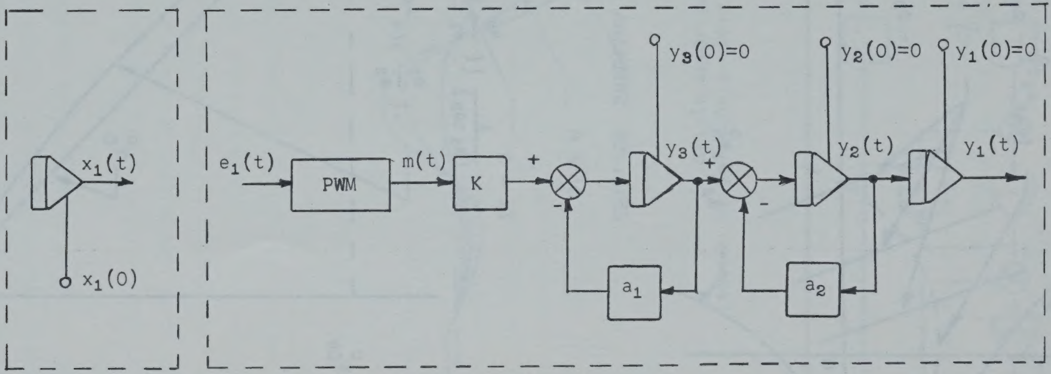

$x_1(t)$
$x_1(0)$
$e_1(t)$
PWM
$m(t)$
K
$y_3(0)=0$
$y_2(0)=0$
$y_1(0)=0$
$y_3(t)$
$y_2(t)$
$y_1(t)$
a_1
a_2
+
-

Fig. 4. Representation of a third-order process

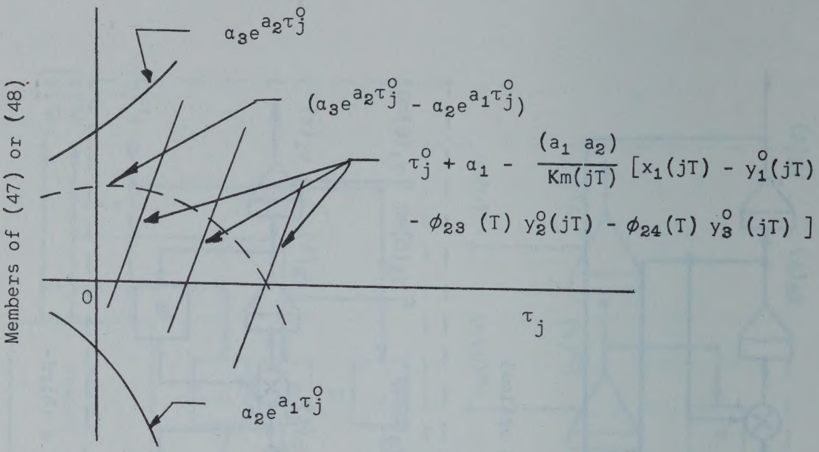

$\alpha_3 e^{a_2 \tau_j^o}$
$(\alpha_3 e^{a_2 \tau_j^o} - \alpha_2 e^{a_1 \tau_j^o})$
$\tau_j^o + \alpha_1 - \frac{(a_1\ a_2)}{Km(jT)} [x_1(jT) - y_1^o(jT)$
$- \phi_{23}(T)\ y_2^o(jT) - \phi_{24}(T)\ y_3^o(jT)\ ]$
Members of (47) or (48)
0
τ_j
$\alpha_2 e^{a_1 \tau_j^o}$

Fig. 5. Solution for optimum pulse width

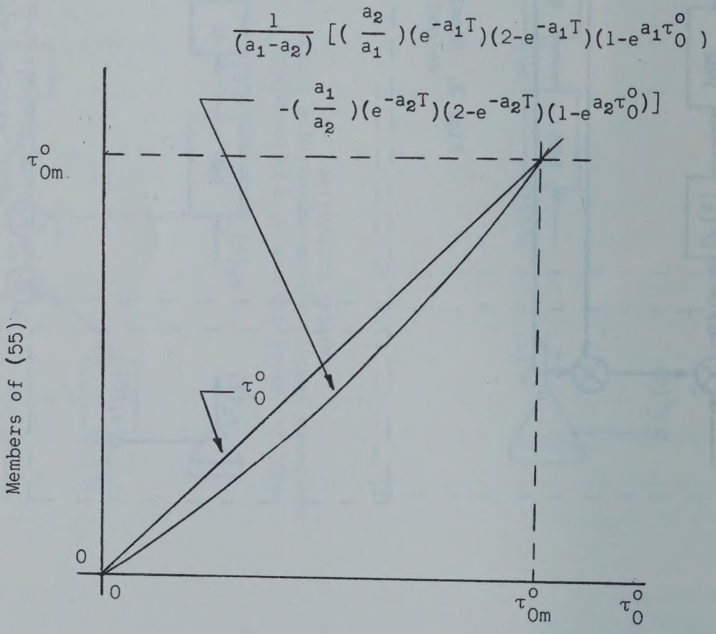


Fig. 6. Determination of the maximum allowable value of initial pulse width

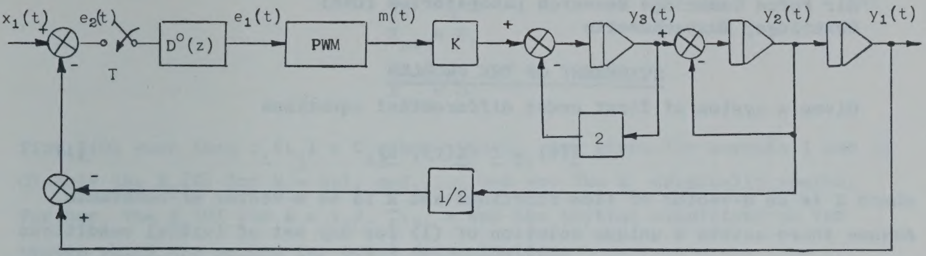

$x_1(t)$
+
$e_2(t)$
T
$D^O(z)$
$e_1(t)$
PWM
m(t)
K
+
$y_3(t)$
+
$y_2(t)$
$y_1(t)$
-
-
-
2
-
1/2
+

Fig. 7. A block diagram of an optimal third-order system

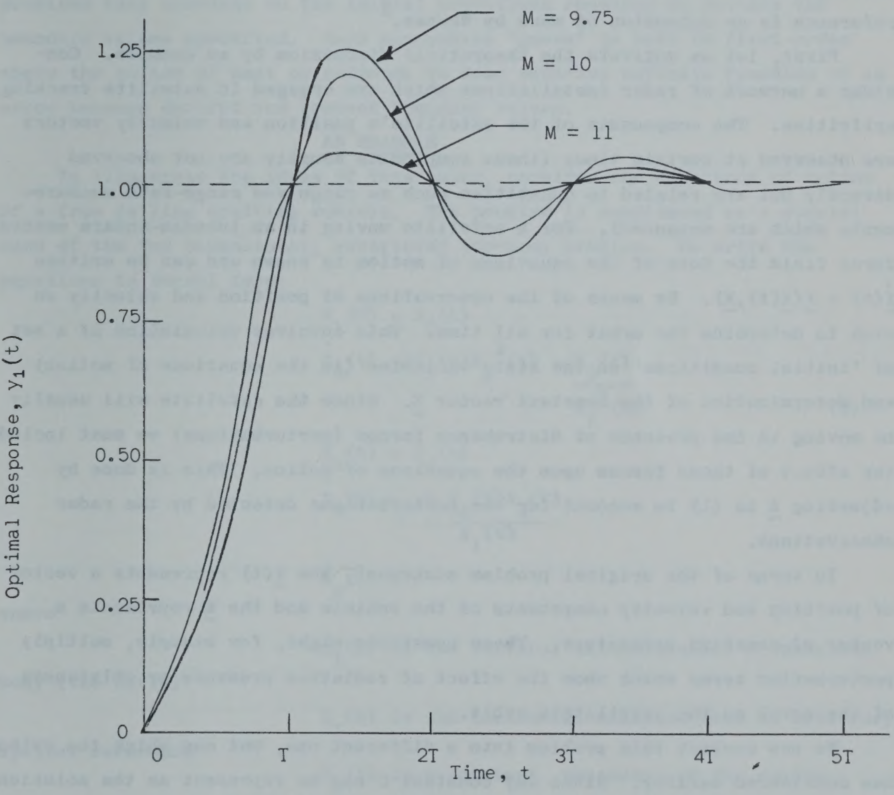


Fig. 8. Response of the optimum system, with M as a parameter

ORBIT DETERMINATION USING ARBITRARILY SPACED DATA

STEPHEN J. KAHNE
Data Sciences Laboratory
Air Force Cambridge Research Laboratories (OAR)
Cambridge, Massachusetts

STATEMENT OF THE PROBLEM

Given a system of first order differential equations

$$\dot{\underline{X}}(t) = \underline{f}\,(\underline{X}(t),\ \underline{K}) \tag{1}$$

where $\underline{X}$ is an n-vector of time functions and $\underline{K}$ is an m-vector of constants. Assume there exists a unique solution of (1) for any set of initial conditions $\underline{X}(0)$ and any $\underline{K}$. If certain components of $\underline{X}(t)$ are given at the times specified below, determine $\underline{X}(t)$ and $\underline{K}$. There are m + n given data points $X_i(t_j)$.

DISCUSSION

We solve this problem by using a scheme due to Cicala,[1] and a method for solving n-point boundary value problems discussed by Kahne.[2] The latter reference is an extension of work by Warner.[3]

First, let us motivate the theoretical discussion by an example. Consider a network of radar installations which are engaged in satellite tracking activities. The components of the satellite's position and velocity vectors are observed at certain times (these components usually are not observed directly but are related to quantities such as range and range-rate measurements which are measured). For a satellite moving in an inverse-square central force field the form of the equations of motion is known and can be written $\dot{\underline{X}}(t) = \underline{f}(\underline{X}(t),\underline{K})$. By means of the observations of position and velocity we wish to determine the orbit for all time. This involves calculation of a set of 'initial conditions' on the state variables (in the equations of motion) and determination of the constant vector $\underline{K}$. Since the satellite will usually be moving in the presence of distrubance forces (perturbations) we must include the effect of these forces upon the equations of motion. This is done by adjusting $\underline{K}$ in (1) to account for the perturbations detected by the radar observations.

In terms of the original problem statement, the $\underline{X}(t)$ represents a vector of position and velocity components of the vehicle and the $\underline{K}$ represents a vector of constant parameters. These constants might, for example, multiply perturbation terms which show the effect of radiation pressure or oblateness of the earth on the satellite's orbit.

We now convert this problem into a different one, but one which the author has considered earlier. Since any constant C can be represent as the solution to the following differential equation

$$\begin{aligned} \dot{X}(t) &= 0 \\ X(0) &= C \end{aligned} \tag{2}$$

we may rewrite the original problem as follows:

Given

$$\begin{aligned} \dot X_1 &= f_1(\underline{X}) \\ \dot X_2 &= f_2(\underline{X}) \\ &\;\vdots \\ \dot X_n &= f_n(\underline{X}) \\ \dot X_{n+1} &= 0 \\ &\;\vdots \\ \dot X_{n+m} &= 0 \end{aligned} \tag{3}$$

find $\underline{X}(0)$ such that $X_i(t_j) = C_{ij}$ where the C_{ij} are given for certain i and j. Clearly the $X_k(0)$ for k = n+1, n+2,..., n+m are the K_i originally sought. Further, the $X_k(0)$ for k = 1,2, ..., n are the initial conditions on the system which are needed for orbit determination.

One solution to this problem as stated above has been described by the author in reference 2. The method shown there involves the expansion of the exact values of unknown initial conditions about 'guesses' for these initial conditions. The 'guesses' are successively improved and for certain classes of problems they converge to the initial conditions required to satisfy the boundary values specified. Each successive 'guess' is best to first order where the notion of best corresponds to some positive definite function of an error between desired and present boundary values.

AN EXAMPLE

To illustrate the ideas of this paper, consider the equations of motion of a free falling orbiting vehicle. The problem is considered as a special case of the two dimensional, restricted two-body problem. We write the equations in normal form:

$$\begin{aligned} \dot X_1(t) &= X_2(t) \\ \dot X_2(t) &= X_1(t)X_2^{\,2}(t) - \frac{X_5(t)}{X_1^{\,2}(t)} \\ \dot X_3(t) &= X_4(t) \\ \dot X_4(t) &= \frac{-2\, X_2(t)X_4(t)}{X_1(t)} \\ \dot X_5(t) &= 0 \end{aligned} \tag{4}$$

where

$X_1(t)$ is the vertical distance above a spherical body (the earth)

$X_3(t)$ is the horizontal distance from an arbitrary spatial reference

$X_5(t)$ is a constant parameter of the system

Initially the horizontal and vertical position of the vehicle is noted and subsequent (radar) measurements yield exact values of X_1 and X_3 at later times. The technique outlined above and in reference 2 generates a set of initial conditions which determines $\underline{X}(t)$ to minimize the error between the boundary conditions and the components of $\underline{X}$ at appropriate times. This error is computed as the sum of $|X_i(t_j) - C_{ij}|$ for the appropriate i and j. Specifically, the following data is used:

$$X_1(0) = 10.000$$
$$X_3(0) = 0.000$$
$$X_1(2) = 22.350$$
$$X_3(2) = 1.108$$
$$X_1(4) = 41.200$$

The criterion for successful convergence is that the error be less than .005. A table showing the results of the iterations is given next. It might be noted that the original 'guesses' required to start the iterative procedure were selected by a rule which is in no way related to the specific problem. (An example of such a rule would be to select the 'guesses' to be 1.00, 2.00, 3.00, etc. for all problems.)

Iteration	$X_1(0)$	$X_1(2)$	$X_1(4)$	$X_3(0)$	$X_3(2)$
1	10.000	22.902	41.398	0.000	1.014
2	10.000	22.902	41.398	0.000	1.014
3	10.000	22.964	41.452	0.000	1.063
4	10.000	22.935	41.319	0.000	1.109
5	10.000	22.321	41.158	0.000	1.108
6	10.000	22.321	41.158	0.000	1.108
7	10.000	22.350	41.218	0.000	1.108
8	10.000	22.350	41.218	0.000	1.108

The initial conditions corresponding to the eighth iteration (and consequently the solution to the problem) follow:

$$X_1(0) = 10.000$$
$$X_2(0) = -0.011$$
$$X_3(0) = 0.000$$
$$X_4(0) = 1.000$$
$$X_5(0) = 0.000$$

Approximately five minutes of (IBM 7094) computer time was required to obtain this solution. As is observed with many iterative schemes, the first iteration produces a set of initial conditions quite close to the correct values. Later iterative steps refine the results of the first. Although certain of the successive iterations appear to produce no change in the functions, the numbers in the table have been rounded, and consequently the improvements are not shown.

CONCLUDING REMARKS

Convergence of the method is not assured and depends upon both the initial estimates of the unknown initial conditions and the extent of the nonlinearity in the equations. Knowledge about probable correct values of the unknown initial boundary values should be used when available. Since the method only uses first order corrections to the initial conditions, even 'good' initial estimates may be inadequate to assure convergence. To improve convergence properties of the technique a second order version has been developed by the author and has proved valuable for 'more difficult' problems. The results of this work will be reported at a later date.

The results of a method such as the one used here are no better than the integration technique employed. In this work a variable step-size method of numerical integration due to A. Nordsieck, formerly at the university of Illinois, was used. This method was coded in FORTRAN by P. Ponzo at the Coordinated Science Laboratory, University of Illinois.

In conclusion it should be noted that efficient solution of multi-point boundary value problems is of increasing importance in the field of system theory. The ideas given here are steps toward efficient numerical solutions.

REFERENCES

1. Kelley, H. J. "Method of Gradients" in Optimization Techniques edited by Leitmann, Academic Press, 1962.
2. S. J. Kahne, "Note on Two Point Boundary Value Problems," IEEE Transactions On Automatic Control, Vol. AC - 8 No. 2, July, 1963.
3. L. Fox, Numerical Solution of Ordinary and Partial Differential Equations, Addison Wesley, Pg 64-66, 1962.
4. R. Bellman, H. Kagiwada, R. Kalaba, "Quasilinearization, System Identification, and Prediction " RAND Memorandum RM-3812-PR, August 1963
5. S. J. Kahne, "On Direct Fixed - Time Optimization of Invertible Systems", R-162 Coordinated Science, University of Illinois, June, 1963.

IDENTIFICATION AND OPTIMIZATION OF MULTIPARAMETER SYSTEMS USING CORRELATION TECHNIQUES

KUMPATI S. NARENDRA and LYLE E. McBRIDE, JR.
Division of Engineering and Applied Physics
Harvard University
Cambridge, Massachusetts

ABSTRACT

A method is described for the identification and optimization of multivariable systems by altering linear parameters to reduce a mean-square performance criterion. The change in each parameter is determined by an error gradient in parameter space computed by cross-correlation techniques which are independent of signal spectra.

The application of this method to several multivariable system configurations is illustrated; the results of analog computer simulation are included to demonstrate the practicality of such systems.

I INTRODUCTION

In previous reports [1] several systems were proposed which employ cross-correlation of measurable signals to compute the gradient of an average error function with respect to a set of linear system parameters. The gradient thus computed is used to alter the parameters along a path of steepest descent toward a point of minimum mean-square error. Examples were given illustrating the application of this technique to the identification of a single-input system and to the self-optimization of feedback control systems.

Identification was accomplished by comparing the unknown system to a model consisting of a sum of fixed transfer functions followed by variable gains. While such a model can in theory represent any transfer characteristic if the component functions constitute a complete set, in practice the number of terms necessary for a reasonable approximation may be large. A more accurate and less complex model could be obtained by altering the poles as well as the gains of the component transfer functions. Configurations similar to that used in the state-vector method of time domain analysis are suitable models with adjustable poles for systems with multiple inputs and outputs.

The application of correlation gradient-computing techniques to multivariable problems is the object of this paper. Also included are the results of analog computer simulation of self-optimizing systems which indicate the practicality of on-line parameter alteration based these principles.

II SYSTEM REPRESENTATION

Before a gradient in parameter space can be defined, a system representation or configuration must be selected to determine the effect of each parameter. A model used to identify an unknown plant should produce approximately the same output when subjected to the same input. Clearly, characteristics of the plant which do not affect its output are unimportant here and so the model need contain no provision for unobservable states. Similarly, since the identification system has access only to information contained in external (input and output) signals, the model will be completely unable to reproduce any output which is not dependent on the past history of the inputs. In a linear device, this statement is equivalent to the equation:

$$y_i(t) = \int_{-\infty}^{t} \sum_{j=1}^{n} h_{ij}(t,\tau)u_j(\tau)d\tau \qquad i=1,\ldots,m. \qquad (1)$$

where y_i is an output, u_j is an input, and $h_{ij}(t,\tau)$ is the response of y_i at time t to an impulse in u_j occurring at time τ. If the system is in addition assumed to be stationary, equation (1) can be replaced by the Fourier transform of its matrix equivalent:

$$\underline{y}(j\omega) = \underline{H}(j\omega)\underline{u}(j\omega) \qquad (2)$$

where $\underline{u}$ and $\underline{y}$ are input and output vectors and $\underline{H}$ is the matrix of transfer functions.

Thus the model can, without restricting its generality, be represented by a matrix of transfer functions. Gilbert [2] has shown that such a matrix defines only the completely controllable, completely observable part of a system, and that its state vector representation is not unique. Therefore identification does not imply a unique determination of the system differential equations, but only a knowledge of transfer characteristics (whether in the time or frequency domain).

STATE VECTOR REPRESENTATION

Although insufficient information is available for a unique determination of the state vector representation of a system, the imposition of certain constraints on the form of the matrices results in a convenient set of identification parameters.

An m-input, ℓ-output linear stationary system of order n (n state variables) can be represented by the matrix equations:

$$\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{D}\,\underline{u} \qquad \underline{y} = \underline{H}\,\underline{x} \tag{3}$$

where $\underline{x}$ is the state vector; $\underline{A}$, $\underline{D}$ and $\underline{H}$ are n x n, n x m and ℓ x n constant matrices, respectively; $\underline{u}$ is the input vector; $\underline{y}$ is the output vector.

The transform of equation (3) can be written as:

$$\underline{y}(j\omega) = \underline{H}(j\omega\underline{I} - \underline{A})^{-1}\underline{D}\,\underline{u} \tag{4}$$

where $\underline{I}$ is the identity matrix of order n. It will be assumed that $\underline{A}$ is non-defective (all homogeneous solutions of equation (3) are sums of exponentials). In this case, a similarity transformation can always be found which diagonalizes $\underline{A}$ and hence $(j\omega\underline{I} - \underline{A})^{-1}$, and equation (3) can be written as:

$$\underline{y}(j\omega) = \hat{\underline{H}}\,\underline{\Lambda}\,\hat{\underline{D}}\,\underline{u} \tag{5}$$

where $\underline{\Lambda} = \underline{S}(j\omega\underline{I} - \underline{A})^{-1}\underline{S}^{-1}$ and is diagonal, $\hat{\underline{H}} = \underline{H}\,\underline{S}^{-1}$, $\hat{\underline{D}} = \underline{S}\,\underline{D}$. It is thus clear that since a change in $\underline{H}$ and $\underline{D}$ can reduce $\underline{A}$ to diagonal form without altering the transfer characteristics, only n elements of $\underline{A}$ need be chosen to fix the eigenvalues or natural frequencies of the system; any other variation in the transfer matrix can be obtained by altering $\underline{H}$ and $\underline{D}$ only.

Furthermore, not all elements in $\underline{H}$ and $\underline{D}$ are necessary to determine the transfer matrix. This can be seen from the subscript form of (5):

$$y_i = \sum_{k,j} h_{ik}\lambda_{kk}d_{kj}u_j \tag{6}$$

Equation (6) is unaltered if each row of $\underline{D}$ is divided by a constant, and the corresponding column of $\underline{H}$ is multiplied by the same constant, since

$$\alpha_k h_{ik}\lambda_{kk}\frac{d_{kj}}{\alpha_k} = \hat{h}_{ik}\lambda_{kk}\hat{d}_{kj} \tag{7}$$

However, since α_k is arbitrary it may be chosen to give any desired value to one column of $\hat{d}$. For example if $\alpha_k = d_{k1}$ then $\hat{d}_{k1} = 1$ for all k.

To identify the most general possible transfer matrix which can be represented by a system of order n, it is sufficient to determine n elements of the matrix $\underline{A}$ which fix its eigenvalues, ℓ x n elements of $\underline{H}$, and n x (m-1) elements of $\underline{D}$, a total of n x (ℓ+m) elements. For convenience in identification it will be assumed in subsequent sections that each element in the first column of $\underline{D}$ is always unity, and that $\underline{A}$ takes on one of the two standard forms of Fig. 1. Form (a) reduces the model to a set of n decoupled first order systems, preceded and followed by a matrix of gains; it simplifies the model construction but is limited to the simulation of systems which have real eigenvalues. Form (b) represents $\frac{n}{2}$ independent second order systems and is capable of representing systems with any combination of real and complex conjugate eigenvalues. Since the use of form (b) complicates the gradient computer, it will not be considered in this paper. Systems having complex eigenvalues are discussed in reference [3].

III IDENTIFICATION OF MULTIVARIABLE SYSTEMS

The general approach to be followed in identification and optimization is the same as that of reference [1]: the partial derivative of a mean square error criterion with respect to a given linear parameter is found to be proportional to the mean of the product (or cross-correlation) of two signals which can be obtained as outputs from the system itself or from suitable models. The rate of change of each parameter is then made proportional to the negative of the corresponding partial derivative so that the system follows a path of steepest descent in parameter space to the optimum points where the gradient is identically zero. It is assumed that this alteration of parameters is sufficiently slow so that the system can be considered linear and stationary for the purpose of computing the error. It is also assumed that the input signals are nearly stationary or periodic over the time required for the system to approach an optimum. If these assumptions hold, the finite time average necessary in practice is approximately equal to the infinite time average used in the following derivations, and the steepest descent procedure can be made to converge for either random or periodic deterministic inputs. (A more detailed discussion of these assumptions is given in [1].)

The time average of the product of two bounded continuous real functions of time x(t) and y(t) is defined as:

$$\overline{x(t)y(t)} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t)y(t)dt = \lim_{T \to \infty} \frac{1}{2T} \int_{-\infty}^{\infty} x_T(t)y_T(t)dt \qquad (8)$$

The Fourier transform of the time limited function can be defined as:

$$x_T(j\omega) = \int_{-T}^{T} x(t)e^{j\omega t}\, dt = \int_{-\infty}^{\infty} x_T(t)e^{j\omega t}\, dt \qquad (9)$$

Then by Parseval's theorem:

$$\begin{aligned}\overline{x(t)y(t)} &= \lim_{T \to \infty} \frac{1}{2T} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} x_T(j\omega)y_T^*(j\omega)d\omega\right] \\ &= \lim_{T \to \infty} \frac{1}{2T} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} x_T^*(j\omega)y_T(j\omega)d\omega\right]\end{aligned} \qquad (10)$$

where $y_T^*(j\omega)$ is the complex conjugate of $y_T(j\omega)$.

In order to simplify the notation, the subscript T, the factors $\frac{1}{2T}$ and $\frac{1}{2\pi}$, and the indication of the limiting process will be omitted and equation (10) written as:

$$\overline{x(t)y(t)} = \int_{-\infty}^{\infty} x(j\omega)y^*(j\omega)d\omega \qquad (11)$$

All equations of the form of Eq. (11) are to be interpreted in the form of Eq. (10).

To identify an unknown plant transfer matrix $\underline{P}(s)$, it is assumed that a set of stationary or periodic inputs is applied simultaneously to the plant and to a model, as shown in Fig. 2, where $\underline{d}(t)$ is the vector of actual plant outputs, and $\underline{e}(t) = \underline{d}(t) - \underline{y}(t)$ is the difference vector between the plant and model outputs.

If the identifying model of Fig. 2 is described by equations (3) and (4), (Fig. 3) and the appropriate error function to be minimized is the average value of the Euclidean norm of the error, $\overline{\underline{e}'\underline{e}}$, the following expression for the error gradients results:

$$\frac{\partial \overline{\underline{e}'\underline{e}}}{\partial h_{ij}} = -2\int_{-\infty}^{\infty} \underline{e}^{*\prime}(j\omega)\,\frac{\partial \underline{H}\,\underline{x}(j\omega)}{\partial h_{ij}}\,d\omega = -2\int_{-\infty}^{\infty} e_i^{*}(j\omega)x_j(j\omega)d\omega$$

$$= -\overline{2e_i(t)x_j(t)} \qquad (12)$$

Since e_i and x_j are both measurable signals the gradient with respect to the elements of $\underline{H}$ can be computed directly from this expression.

The derivative with respect to the elements of $\underline{A}$ is:

$$\frac{\partial \overline{\underline{e}'\underline{e}}}{\partial a_{ij}} = -2\int_{-\infty}^{\infty} \underline{e}^{*\prime}\,\frac{\partial \underline{H}(j\omega\underline{I} - \underline{A})^{-1}\underline{D}\,\underline{u}}{\partial a_{ij}}\,d\omega$$

$$(13)$$

$$= -2\int_{-\infty}^{\infty} \underline{e}^{*\prime}\underline{H}(j\omega\underline{I} - \underline{A})^{-1}\,\frac{\partial \underline{A}}{\partial a_{ij}}\,(j\omega\underline{I} - \underline{A})^{-1}\underline{D}\,\underline{u}\,d\omega$$

If the matrix $\underline{\Phi}$ is defined by:

$$\underline{\Phi} = (j\omega\underline{I} - \underline{A})^{-1} \qquad (14)$$

Eq. (13) can be written in subscript notation as:

$$\frac{\partial \overline{\underline{e}'\underline{e}}}{\partial a_{ij}} = -2\int_{-\infty}^{\infty} \sum_{\ell,m,r,s} e_{\ell}^{*}h_{\ell m}\phi_{mi}\phi_{jr}d_{rs}u_s d\omega$$

$$(15)$$

$$= -2\int_{-\infty}^{\infty} \sum_{\ell,m} e^{*}_{\ell}h_{\ell m}\phi_{mi}x_j d\omega$$

Similarly:

$$\frac{\partial \overline{\underline{e}'\underline{e}}}{\partial d_{ij}} = -2\int_{-\infty}^{\infty} \underline{e}^{*\prime}\underline{H}\,\underline{\Phi}\,\frac{\partial \underline{D}}{\partial d_{ij}}\,\underline{u}\,d\omega$$

$$= -2\int_{-\infty}^{\infty} \sum_{\ell,m} e^*_\ell h_{\ell m}\phi_{mi} u_j \, d\omega \tag{16}$$

If the model is restricted to the identification of systems having real roots, the matrix $\underline{A}$ is diagonal (Fig. 1a), $\underline{\Phi}$ is diagonal, and $\phi_{ii} = \frac{1}{s - a_i}$. In this case, equations (15) and (16) can be written:

$$\frac{\partial \overline{\underline{e}'\underline{e}}}{\partial a_i} = -2\int_{-\infty}^{\infty} \sum_{\ell} e^*_\ell h_{\ell i}\left(\frac{1}{j\omega - a_i}\right) x_i \, d\omega$$

$$\frac{\partial \overline{\underline{e}'\underline{e}}}{\partial d_{ij}} = -2\int_{-\ell}^{\ell} \sum_{\ell} e^*_\ell h_{\ell i}\left(\frac{1}{j\omega - a_i}\right) u_j \, d\omega \tag{17}$$

Then the gradient can be computed by the systems shown in Fig. 4, since $\eta^*_i = \sum_\ell e^*_\ell h_{\ell i}$, $\delta_i = \frac{x_i}{j\omega - a_i}$, ${}^j\gamma_i = \frac{u_j}{j\omega - a_i}$ and therefore:

$$\frac{\partial \overline{\underline{e}'\underline{e}}}{\partial a_i} = -\overline{2\delta_i(t)\,\eta_i(t)}, \quad \frac{\partial \overline{\underline{e}'\underline{e}}}{\partial d_{ij}} = -\overline{2\,{}^j\gamma_i(t)\,\eta_i(t)} \tag{18}$$

Note that one dynamic model is required to compute the gradient with respect to the a_i, and several additional models (one for each input variable except the first, since d_{i1} are assumed constant) are required for the derivatives with respect to d_{ij}; since the values of a_i in all models must be altered simultaneously as the system optimizes itself, this configuration is rather cumbersome if very many inputs are present.

This multiplicity of models can be eliminated by the use of an adjoint model (simulated, for example, by a digital computer). If a vector $\underline{\xi}$ is defined by:

$$\xi(j\omega) = \underline{\Phi}'^*\underline{H}'\underline{e} = (-j\omega\underline{I} + \underline{A}')^{-1}\underline{H}'\underline{e} \tag{19}$$

Eqs. (15) and (16) can be written:

$$\frac{\partial \overline{\underline{e}'\underline{e}}}{\partial a_{ij}} = -\overline{2\xi_i(t)x_j(t)}$$

$$\frac{\partial \overline{\underline{e}'\underline{e}}}{\partial d_{ij}} = -2\xi_i(t)u_j(t) \tag{20}$$

requiring only a single model regardless of the form of $\underline{A}$.

Computer simulation: The single-input, single-output two-parameter identification problem was simulated on the analog computer. Since the first column of $\underline{D}$ was to be made up of ones, the single-variable model with a diagonal matrix $\underline{A}$ can be regarded as a combination of parallel networks, $\frac{K_i}{s + \lambda_i}$, whose outputs are summed. Two cases were considered.

1) Fixed poles with adjustable gains: The problem of identifying a second order plant transfer function as $\frac{K_1}{s + 1} + \frac{K_2}{s + 2}$ was simulated; it was found that extremely rapid convergence could be obtained when the plant to be identified contained the same poles as the model. Fig. 5 shows the change in K_1, K_2 and e(t) when plant and model were subjected to a square wave of about 1/4 cycle per second. The plant transfer function was $\frac{10}{(s + 1)(s + 2)}$. The theoretical equilibrium values for K_1 and K_2 for perfect identification are 10 and -10. The value of β employed was 0.05.

Other runs demonstrated the fact that the self-optimizing system reduced its error rapidly even when the plant poles were far removed from those of the identification networks; however, under these conditions the optimum reached varied greatly as a function of the type of input signal applied.

In setting up the self-optimizing system on the computer the averaging operation was omitted, and the $\eta_i e$ product integrated directly to compute K_i; this omission was justified by the fact that with the small values of β used the K's were relatively constant at their average values when equilibrium was reached. Attempts to approximate an averaging operation by the insertion of a first order transfer function (e.g. $\frac{1}{10s + 1}$) showed that the added phase lag due to averaging reduced the stability of the closed parameter-adjustment loop and thus decreased the allowable speed of convergence. It should also be noted that because of the nonlinearity of the correlation operation the stability of parameter-adjustment is decreased when input amplitude is increased; conversely a choice of β which results in rapid convergence with a large input amplitude will produce much slower convergence at lower signal levels.

2) Adjustable pole and gain: The identification of a first order plant as $\frac{K}{s+\lambda}$ was simulated, and the result is shown in Fig. 6. The plant to be identified was in this case $\frac{5}{s+1}$; the initial values of K and λ were 0 and 2 respectively. Although convergence is not as rapid as that in Fig. 5 (where the same input was applied), the system proved capable of determining the best mean square first order approximation; the system of Fig. 5, which is nearly as complex, is limited to an approximation with two fixed poles.

IV SELF-OPTIMIZING CONTROL SYSTEMS

Correlation techniques can be used equally well to optimize parameter values in a closed loop containing a plant whose characteristics cannot be altered. The procedure is essentially the same as for identification; the only difference is the choice of configuration and error criterion.

A) State-variable feedback

One problem in the theory of optimal control involves the computation of the optimal linear combination of state variables which, when fed back into the input of a dynamic system will transfer the state vector from an initial value $\underline{x}^o$ at t_o to the origin at time t_1, while minimizing an error criterion of the form:

$$J = \int_{t_o}^{t_1} \left[\underline{x}' \underline{Q}\, \underline{x} + \underline{u}' \underline{R}\, \underline{u} \right] dt \qquad (19)$$

An infinite time equivalent might be the problem of finding the optimum state-variable feedback matrix for a multivariable servo. The system configuration is that of Fig. 7, while the error criterion becomes:

$$J = \overline{\underline{e}' \underline{Q}\, \underline{e}} + \overline{\underline{u}' \underline{R}\, \underline{u}} \qquad (20)$$

$$\underline{u} = \underline{s} - \underline{G}\, \underline{x} \quad , \quad \underline{e} = \underline{s} - \underline{y} = \underline{s} - \underline{H}\, \underline{x} \qquad (21)$$

where $\underline{Q}$ and $\underline{R}$ are arbitrary symmetric weighting matrices.

Then, since:

$$x(j\omega) = (\underline{I} + \underline{\Phi}(j\omega)\underline{D}\,\underline{G})^{-1}\,\underline{\Phi}(j\omega)\underline{D}\,\underline{s}\ , \tag{22}$$

$$\frac{\partial \underline{x(j\omega)}}{\partial g_{ij}} = -\,(\underline{I} + \underline{\Phi}\,\underline{D}\,\underline{G})^{-1}\,\underline{\Phi}\,\underline{D}\,\frac{\partial \underline{G}}{\partial g_{ij}}\,(\underline{I} + \underline{\Phi}\,\underline{D}\,\underline{G})^{-1}\,\underline{\Phi}\,\underline{D}\,\underline{s} \tag{23}$$

$$= -\,(\underline{I} + \underline{\Phi}\,\underline{D}\,\underline{G})^{-1}\,\underline{\Phi}\,\underline{D}\,\frac{\partial \underline{G}}{\partial g_{ij}}\,\underline{x} \tag{24}$$

the partial derivative of the criterion is:

$$\begin{aligned}\frac{\partial J}{\partial g_{ij}} &= 2\int_{-\infty}^{\infty}\left[\underline{e}'^{*}(j\omega)\underline{Q}\,\frac{\partial \underline{e}(j\omega)}{\partial g_{ij}} + \underline{u}'^{*}(j\omega)\underline{R}\,\frac{\partial \underline{u}(j\omega)}{\partial g_{ij}}\right]d\omega \\ &= 2\int_{-\infty}^{\infty}\left[\underline{e}'^{*}\underline{Q}\,\underline{H} + \underline{u}'^{*}\underline{R}\,\underline{G}\right](\underline{I} + \underline{\Phi}\,\underline{D}\,\underline{G})^{-1}\underline{\Phi}\,\underline{D}\,\frac{\partial \underline{G}}{\partial g_{ij}}\,\underline{x}\,d\omega\end{aligned} \tag{25}$$

Since $\frac{\partial \underline{G}}{\partial g_{ij}}$ is a matrix with all elements zero, except the (i,j)th which is unity, equation (25) can be written:

$$\frac{\partial J}{\partial g_{ij}} = 2\int_{-\infty}^{\infty} \eta_i^{*} x_j\,d\omega = 2\,\overline{\eta_i x_j} \tag{26}$$

where η is the output of the adjoint system of Figure 8,

$$\eta = \underline{D}'\underline{\Phi}'^{*}(\underline{I} + \underline{G}'\underline{D}'\underline{\Phi}'^{*})^{-1}(\underline{H}'\underline{Q}\,\underline{e} + \underline{G}'\underline{R}\,\underline{u}) \tag{27}$$

The gradient $\frac{\partial J}{\partial g_{ij}}$ can also be obtained by using a number of models of the system itself (as in Fig. 4) instead of the adjoint system of Fig. 8 if it is necessary to use analog simulation. Once the product $\eta_i x_j$ is available, a set of integrating parameter controllers will cause a continual reduction in J until the optimal system is attained. The optimum parameter values here depend, of course, on the input signal spectrum, and for convergence the system requires a stationary or periodic input.

Since no restruction has been made on the transfer matrix $\underline{\Phi}$(s) requiring its outputs to be state variables, the same expressions are valid for any plant whose output vector $\underline{x}$ is to be fed back by an optimum matrix of constants (in this case $\underline{H}$ is the unit matrix).

B) An optimal dynamic controller.

In the case where the outputs available from the plant to be controlled are not state variables, it may be desirable to include dynamics in the closed loop controller. A possible configuration is shown in Fig. 9.

If the controller is assumed to have the form of Fig. 3 with any desired number of state variables, the transforms of the error $\underline{e}$ and the input $\underline{u}$ are given by:

$$\underline{e}(j\omega) = [\underline{I} + \underline{P}(j\omega)\underline{H}(j\omega\underline{I} - \underline{A})^{-1}\underline{D}]^{-1}\underline{s} \tag{28}$$

$$\underline{u}(j\omega) = \underline{H}(j\omega\underline{I} - \underline{A})^{-1}\underline{D}\ \underline{e}(j\omega) = \underline{H}\ \underline{\Phi}(j\omega)\underline{D}\ \underline{e}(j\omega) = \underline{H}\ \underline{x}$$

$$\frac{\partial \underline{e}}{\partial h_{ij}} = -[\underline{I} + \underline{P}\ \underline{H}\ \underline{\Phi}\ \underline{D}]^{-1}\underline{P}\frac{\partial \underline{H}}{\partial h_{ij}}\ \underline{\Phi}\ \underline{D}[\underline{I} + \underline{P}\ \underline{H}\ \underline{\Phi}\ \underline{D}]^{-1}\underline{s} \tag{29}$$

$$= -[\underline{I} + \underline{P}\ \underline{H}\ \underline{\Phi}\ \underline{D}]^{-1}\ \underline{P}\ \frac{\partial \underline{H}}{\partial h_{ij}}\ \underline{x}$$

$$\frac{\partial \underline{u}}{\partial h_{ij}} = \frac{\partial H}{\partial h_{ij}}\ \underline{x} + \underline{H}\ \underline{\Phi}\ \underline{D}\frac{\partial e}{\partial h_{ij}} \tag{30}$$

$$\frac{\partial \underline{e}}{\partial a_{ij}} = -[\underline{I} + \underline{P}\ \underline{H}\ \underline{\Phi}\ \underline{D}]^{-1}\underline{P}\ \underline{H}\ \underline{\Phi}\frac{\partial A}{\partial a_{ij}}\ \underline{\Phi}\ \underline{D}[\underline{I} + \underline{P}\ \underline{H}\ \underline{\Phi}\ \underline{D}]^{-1}\underline{s} \tag{31}$$

$$= -[\underline{I} + \underline{P}\ \underline{H}\ \underline{\Phi}\ \underline{D}]^{-1}\underline{P}\ \underline{H}\ \underline{\Phi}\frac{\partial \underline{A}}{\partial a_{ij}}\ \underline{x}$$

$$\frac{d\underline{u}}{\partial a_{ij}} = \underline{H}\ \underline{\Phi}\frac{\partial A}{\partial a_{ij}}\ \underline{x} + \underline{H}\ \underline{\Phi}\ \underline{D}\ \frac{\partial e}{\partial a_{ij}} \tag{32}$$

$$\frac{\partial \underline{e}}{\partial d_{ij}} = -[\underline{I} + \underline{P}\ \underline{H}\ \underline{\Phi}\ \underline{D}]^{-1}\underline{P}\ \underline{H}\ \underline{\Phi}\ \frac{\partial \underline{D}}{\partial d_{ij}}\ \underline{e} \tag{33}$$

and

$$\frac{\partial \underline{u}}{\partial d_{ij}} = \underline{H}\ \underline{\Phi}\ \frac{\partial \underline{D}}{\partial d_{ij}}\ \underline{e} + \underline{H}\ \underline{\Phi}\ \underline{D}\ \frac{\partial \underline{e}}{\partial d_{ij}} \tag{34}$$

Assuming that the criterion to be minimized is again given by Eq. (20), Eqs. (28) through (34) lead to the following expressions for the derivatives:

$$\frac{\partial J}{\partial h_{ij}} = -\ 2\int_{-\infty}^{\infty} [\underline{e}'{*}\underline{Q}\ \underline{P} - \underline{u}'{*}\underline{R}][\underline{I} + \underline{H}\ \underline{\Phi}\ \underline{D}\ \underline{P}]^{-1}\ \frac{\partial \underline{H}}{\partial h_{ij}}\ \underline{x} \tag{35}$$

$$= -\ 2\ \overline{\eta_i x_j}$$

$$\frac{\partial J}{\partial a_{ij}} = -\ 2\int_{-\infty}^{\infty} [\underline{e}'{*}\underline{Q}\ \underline{P} - \underline{u}'{*}\ \underline{R}]\underline{H}\ \underline{\Phi}[\underline{I} + \underline{D}\ \underline{P}\ \underline{H}\ \Phi]^{-1}\ \frac{\partial \underline{A}}{\partial a_{ij}}\ \underline{x} \tag{36}$$

$$= -2\,\overline{\xi_i x_j}$$

$$\frac{\partial J}{\partial d_{ij}} = -2\int_{-\infty}^{\infty} [\underline{e}'^{*}\underline{Q}\ \underline{P} - \underline{u}'^{*}\underline{R}]\underline{H}\ \underline{\Phi}[\underline{I} + \underline{D}\ \underline{P}\ \underline{H}\ \underline{\Phi}]^{-1} \frac{\partial \underline{D}}{\partial d_{ij}} \underline{e} \qquad (37)$$

$$= -2\,\overline{\xi_i e_j}$$

where $\underline{\xi}$ and η are the outputs of the adjoint model shown in Fig. 10. It is not necessary, of course, to compute these derivatives for more than $n(\ell + m)$ parameters chosen according to the requirements in section II.

Computer simulation: In the single variable case, the controller again is reduced to the form of $\Sigma K_i C_i(s)$. If the $C_i(s)$ are considered fixed (non-adjustable $\underline{A}$ matrix) the parameter computation is simplified but the system retains some of the benefits of a dynamic controller. A commonly used example of such a unit is the so-called three-term chemical process controller whose transfer function is:

$$G(s) = K_1 + \frac{K_2}{s} + sK_3. \qquad (38)$$

where K_1, K_2 and K_3 are manually adjustable for optimum average performance. Then if the error to be minimized is simply $\overline{e^2}$, and a real rather than adjoint model is used, the gradient is given by [1]:

$$\frac{\partial \overline{e^2}}{\partial K_i} = -2\int_{-\infty}^{\infty} e^{*}[1 + P \sum_j K_j C_j(j\omega)]^{-1} PC_i e\, d\omega \qquad (39)$$

$$= -2\,\overline{e(t)\xi_i(t)}$$

where $\xi_i(t)$ is the output of a system model whose input is $e(t)$ (Fig. 11).

This single-variable self-optimizing system was simulated on the analog computer, using the steepest-descent gain-adjusting device already described. The transfer function components used were those of a two-term process controller, K_1, and $\frac{K_2}{s}$. The plant was a third order system whose transfer function was approximately $\frac{1}{(s + 1)(s + 2)(s + 1/2)}$, the input signal was $s(t) \approx (10 \cos 0.9t + 3)$ and $\beta = 0.01$. Fig. 12 shows the error (input minus output) and the two gains for this self-optimizing system when the initial values of K_1 and K_2 were in the unstable region.

The initially oscillatory error is accompanied by a rapid reduction in both parameter values until the oscillations disappear and the only error component is at the input frequency, though reduced by a factor of almost 20. At the end of the 150 seconds the values of K_1 and K_2 were nearly constant, indicating that the gradient of the mean square error was very small.

V CONCLUSION

The examples presented here are an indication that the self-optimization and identification methods of [1] can be applied to complex multivariable problems. For systems with large numbers of inputs and outputs the models required are probably best-suited to digital computer simulation, since real-time self-optimization requires a prohibitive amount of equipment in such cases.

The analog computer results show, however, that for relatively simple systems direct self-optimization or identification is practical, and that the approximations made regarding infinite time averaging and constancy of parameters can be valid in practice.

The multivariable system configurations discussed should be useful in selecting analog models for transfer matrices; it seems from the variety of these models that a gradient computation method based on signal correlation can be derived for almost any imaginable linear system configuration.

REFERENCES

[1] K. S. Narendra and L. E. McBride, "Multiparameter Self-optimizing Systems Using Correlation Techniques", Cruft Laboratory Technical Report No. 413, June 5, 1963, Harvard University, Cambridge, Mass.

[2] E. G. Gilbert, "Controllability and Observability in Multivariable Control Systems", J. Soc. Indust. Appl. Math. Ser. A: On Control, Vol. 1, No. 2 (1963) pp. 128-151.

[3] K. S. Narendra and L. E. McBride, "Identification and Optimization of Multivariable Systems Using Correlation Techniques", Cruft Laboratory Technical Report No. 418, August 1, 1963, Harvard University, Cambridge, Mass.

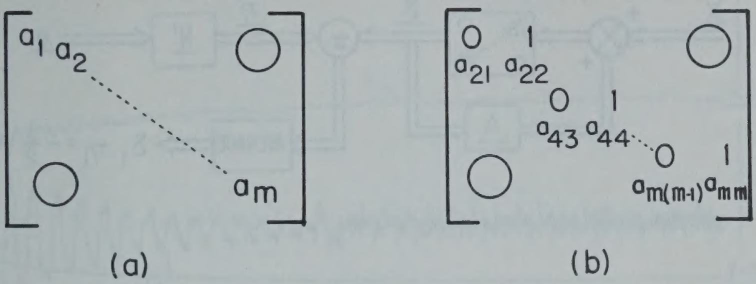

a_1 a_2
a_m
0
0 1
a_{21} a_{22}
0 1
a_{43} a_{44}
0 1
$a_{m(m-1)}$ a_{mm}
0
(a)
(b)

FIGURE 1

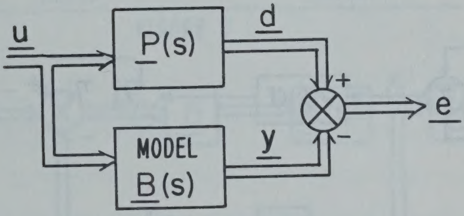

u
P(s)
d
+
e
MODEL
B(s)
y
−

FIGURE 2

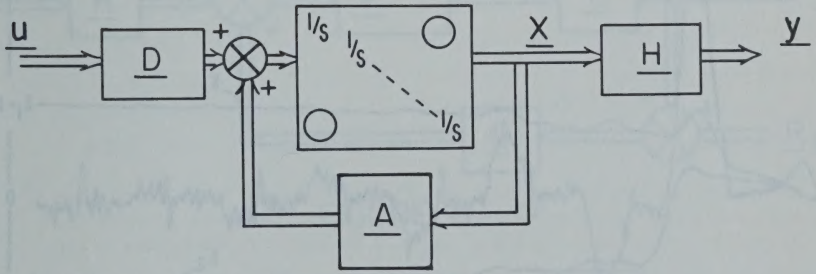

u
D
+
+
1/s
1/s
1/s
0
0
X
H
y
A

FIGURE 3

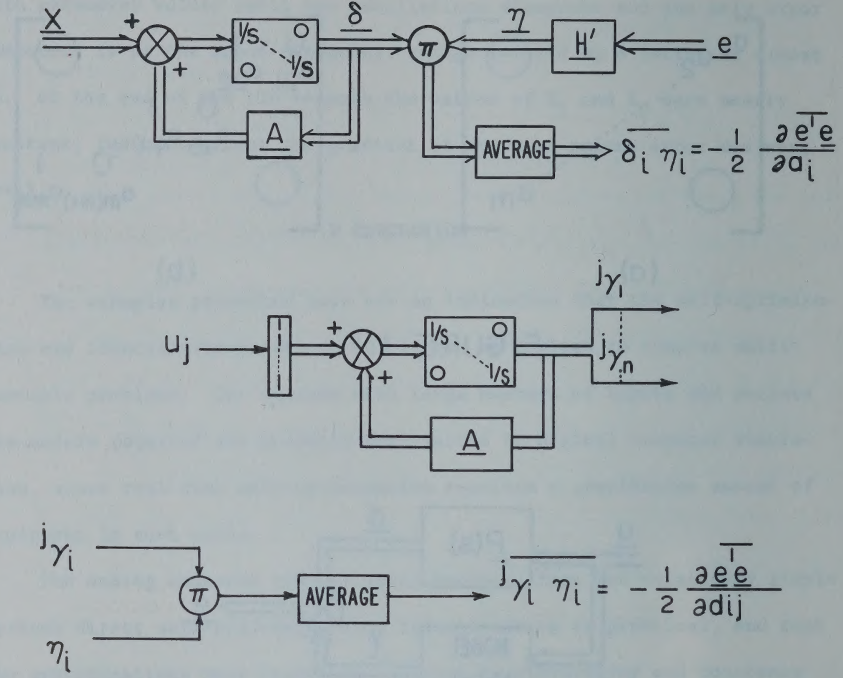

X
+
+
1/S
1/S
δ
π
η
H'
e
A
AVERAGE
$\overline{\delta_i \eta_i} = -\frac{1}{2} \frac{\partial \overline{e^T e}}{\partial a_i}$
u_j
+
+
1/S
1/S
A
$^j\gamma_1$
$^j\gamma_n$
$^j\gamma_i$
π
AVERAGE
η_i
$\overline{{}^j\gamma_i \, \eta_i} = -\frac{1}{2} \frac{\partial \overline{e^T e}}{\partial d_{ij}}$

FIGURE 4

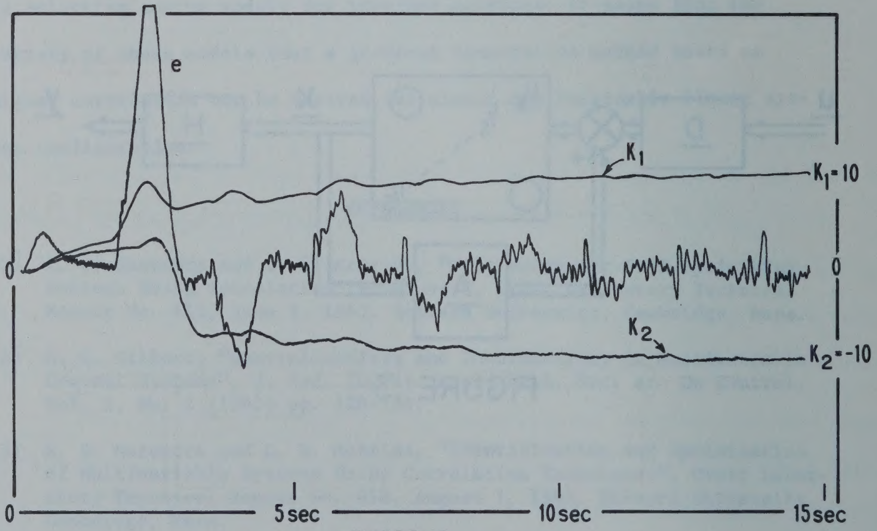


FIGURE 5

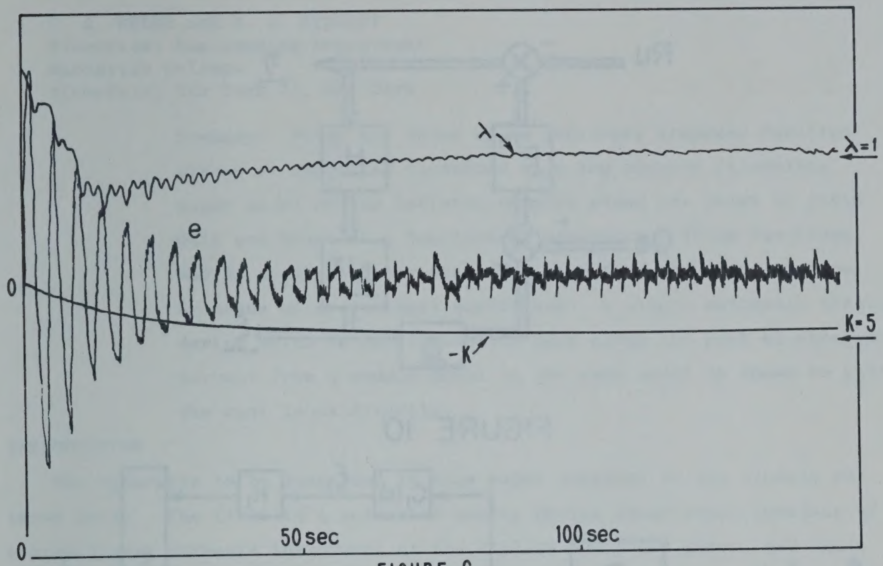


FIGURE 6

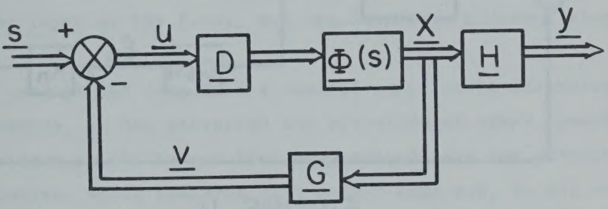

s
+
u
D
Φ(s)
x
H
y
−
v
G

FIGURE 7

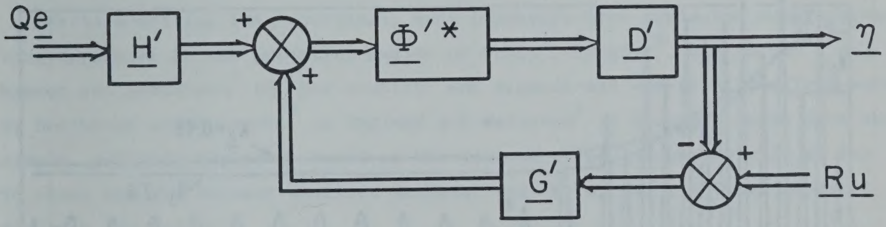


FIGURE 8

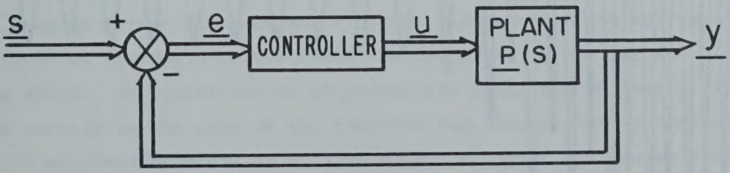

s
+
e
CONTROLLER
u
PLANT
P(S)
y
−

FIGURE 9

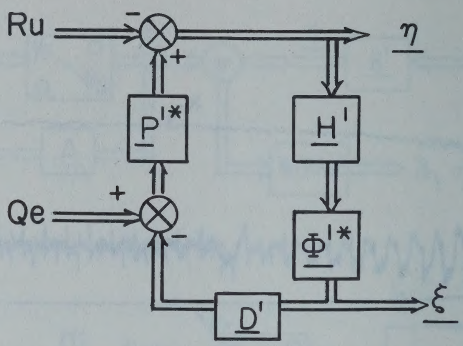

Ru
−
+
η
P'*
H'
Qe
+
−
Φ'*
D'
ξ

FIGURE 10

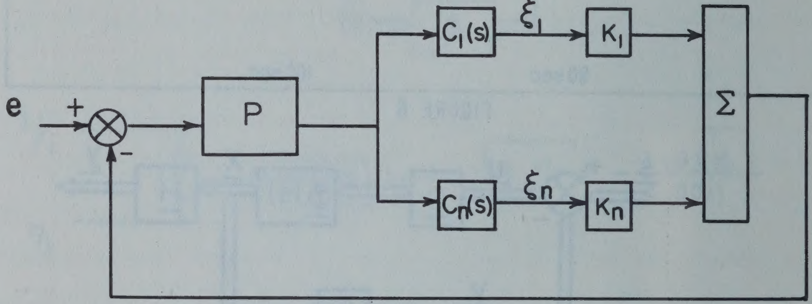

e
+
−
P
C1(s)
ξ1
K1
Cn(s)
ξn
Kn
Σ

FIGURE 11

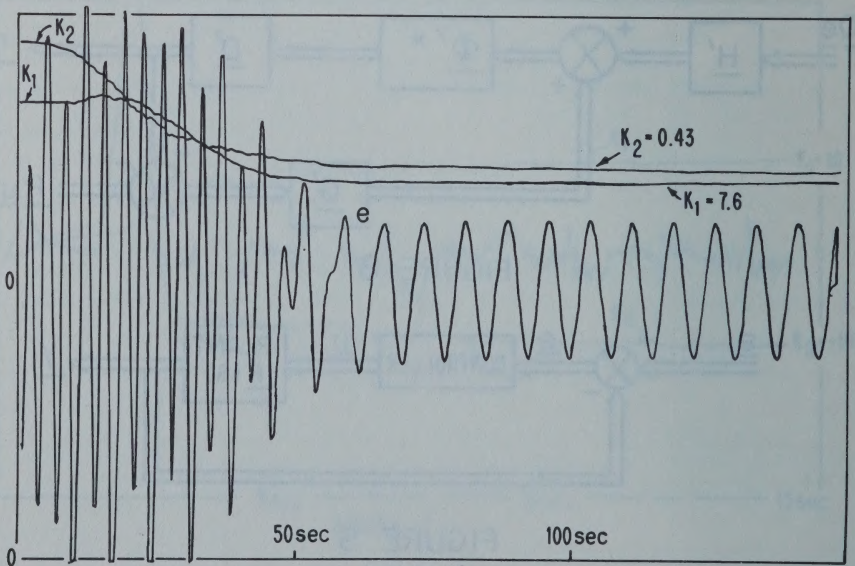


FIGURE 12

THEORETICAL ASPECTS OF AN AUTOMATIC PHASE, GAIN AND ROOT LOCUS PLOTTING DEVICE

C. A. WELSH and B. J. STEWART
Electrical Engineering Department
Manhattan College
Riverdale, New York 71, New York

Summary: Poles and zeros of an arbitrary transfer function placed at specified locations on a two sheeted teledeltos paper model of the infinite complex plane are shown to yield Gain and Phase as a function of frequency. These functions may be simultaneously displayed on a dual beam oscilloscope, by means of operational amplifiers. A simple automatic tracing device which is constrained to move along the path of steepest descent from a saddle point in the same model is shown to plot the root locus directly.

INTRODUCTION

The apparatus to be described in this paper consists of two closely related units. The first is a potential analog device whose input consists of steady direct currents introduced at the s plane poles and zeros, and whose output is a direct plot of gain vs frequency and phase vs frequency. The second device has the same input as the first, but the output is a direct plot of the root locus.

The pole-zero concept has long held a central position in electrical engineering. More recently, it has attracted the attention of other branches of engineering and science to the extent that many schools are now offering a common introductory course, which features this modern approach, to all engineers in the sophomore year. Liberal use of a plane analysis is made in most upper division courses.

While admitting its importance, many educators have neglected such laboratory apparatus as the potential analog of Scott[1], Huggins[2], Darlington[3], and Hansen and Lundstrom[4] in this country, and significant contributions from abroad by Boothroyd and Westcott[5] in England and Malavard[6] in France. It is felt that simple. reliable apparatus based on the work of these innovators will go far to close the gap between abstract mathematical theory and practical physical analysis and design.

In a practical apparatus, a large circle of electrically conducting paper may be used to represent the entire s plane. Unit currents introduced at the poles and removed at the zeros in the s plane, corresponding to any desired transfer function produce a unique field configuration on the surface of the conducting sheet. The distribution of potentials along the $j\omega$ axis will then correspond exactly to the gain of the transfer function if this potential distribution is explored by means of a large number of point electrodes placed along the $j\omega$ axis, the gain as a function of frequency can be displayed directly on one channel of a dual beam oscilloscope.

Phase change along the $j\omega$ axis is proportional to the potential gradient normal to that axis. A measurement of this quantity is most easily performed by sampling the potential a short distance to the left and right of the axis, and then using three operational amplifiers to perform the arithmetic operations involved in the finite difference expansion and the summation process itself. The phase output may be brought to the second vertical channel of the dual beam oscilloscope, resulting in a direct display of phase as a function of frequency. The apparatus thus described is a complete unit in itself.

The root locus method of Evans is now widely accepted as a procedure for determining the closed loop response of a system from the open loop poles and zeros. Boothroyd[5] has shown that the path of steepest descent from a saddle point in a two dimensional conservative field is actually a branch of the root locus. The entire apparatus described above, with the exception of the $j\omega$ axis probe and the rotary switch, can be used to represent this field. The only additional apparatus needed to "close the loop" is an automatic device which will sense the path of steepest descent, and thus describe the root locus.

THEORY OF THE GAIN AND PHASE PLOTTER

Subject to the restrictions of exponential excitation, the finite, inert, linear, lumped-parameter, bilateral passive system will have a response which is characterized by some ratio of through and across variables. If both are through variables or both across variables, the ratio is called a transfer function. If the ratio is mixed, and the through variable is in the numerator, the system function is called either the transfer admittance or the driving point admittance. If the across variable is in the numerator, the system function is the transfer impedance or the driving point impedance. In all cases, these ratios are analytic functions of a complex variable. More explicitly, they are functions of the complex frequency variable, s. The general ratio of through and across variables will now be referred to as the general system function, abbreviated title, GSF(s), and treated in its factored form,

$$\mathrm{GSF}(s) = \frac{k(s-s_1)\ (s-s_3)\ \ldots\ (s-s_m)}{(s-s_2)\ (s-s_4)\ \ldots\ (s-s_n)} \tag{1}$$

where the subscripted s values are, in general, complex. Expression (1) is not suitable for analog simulation in this form, but if the natural logarithm of both sides is taken, we may write it in a form which is suitable.

$$\ln \mathrm{GSF}(s) = \ln \left|\mathrm{GSF}(s)\right| + j\ \mathrm{Arg}\ \mathrm{GSF}(s) \tag{2}$$

$$\begin{aligned}\mathrm{Re}\left\{\ln \mathrm{GSF}(s)\right\} = \ln k + \ln\left|(s-s_1)\right| + \ln\left|(s-s_3)\right| \ldots + \ln\left|(s-s_m)\right| \\ - \ln\left|(s-s_2)\right| - \ln\left|(s-s_4)\right| \ldots - \ln\left|(s-s_n)\right|\end{aligned} \tag{3}$$

If GSF(s) is a dimensionless ratio (transfer function) the real part is recognized as the gain in nepers, and the imaginary part is the phase angle in radians. For any other ratio of through and across variables, the real part displays the change of admittance or impedance with frequency, and the imaginary part the change of phase angle with frequency. The potential analog representation of GSF(s) is as follows.

A current I, introduced at any point in an (effectively) infinite conducting sheet, will cause a potential at any point a distance r from the current probe given by the expression

$$v = K \ln r_k + C \qquad K = \frac{-I\rho}{2\pi t} \tag{4}$$

where K is a function of I the strength of the current, ρ the unit resistivity of the median and t, the thickness of the conducting sheet and C is an arbitrary reference voltage.

For a number of current probes, the total potential is the summation of the contributions from each one individually and the general expression for the potential due to an arbitrary number of equal current sources and sinks located at the subscripted s points becomes,

$$\begin{aligned} v = K \Big[&\ln \left|(s-s_1)\right| + \ln \left|(s-s_3)\right| \ldots + \ln \left|(s-s_n)\right| \\ &-\ln \left|(s-s_2)\right| - \ln \left|(s-s_4)\right| \ldots - \ln \left|(s-s_m)\right| \Big] + C \end{aligned} \tag{5}$$

where $r_k = (s-s_k)$.

The fact that expression (5) has the same mathematical form as (3) except for a constant and a scaling factor, suggests the use of such a distribution of current sources and sinks to represent the poles and zeros of GSF(s) in an (effectively) infinite conducting sheet which is hereafter referred to as the s plane, with coordinate axes σ and $j\omega$. The potential anywhere with respect to a fixed point, is identified as proportional to $\mathrm{Re}\left\{\ln \mathrm{GSF}(s)\right\}$, and in particular, the potential at any point along the $j\omega$ axis is identified as proportional to $\mathrm{Re}\left\{\ln \mathrm{GSF}(j\omega)\right\}$. A scanning of these potentials by means of a row of fixed probes along the $j\omega$ axis, a commutator and a synchronized x axis sweep enables one to plot the logarithm of the gain of a transfer function vs radian frequency on an xy plotter or on one channel of a dual beam oscilloscope.

Although the analog representation of the gain is quite direct, that of the phase function is somewhat less straightforward. Equation (2) has shown logarithmic gain and phase to be conjugate harmonic functions of the complex frequency variable, s. In the analog, the function which is the harmonic conjugate of the potential is the flux. Physically, the flux to be determined for the representation of the total phase shift from zero frequency to any desired frequency ω_p amounts to the total current crossing the $j\omega$ axis between the points $\omega = 0$

and $\omega = \omega_p$.

Generally, the measurement of any through variable demands a cut, and the insertion of a through variable meter. Since current is a through variable, the direct measurement of the current passing between the two points would involve the impractical scheme of

1. Cutting the conducting paper along the $j\omega$ axis from 0 to ω_n.
2. Closing the cut by a "dense set" of current meters.
3. Summing the current read on each meter to give the total flux.

Instead, an indirect measurement can be made in a very practical manner. Essentially, this means that the direct measurement of the through variable which demands a cut, would be replaced by a measurement of the across variable, potential, which does not demand a cut.

The current density in the neighborhood of a j-axis point is proportional to the potential gradient normal to the j-axis at that point. This potential gradient is in actuality the derivative, $dv/d\sigma$. However, this derivative may be replaced for all practical purposes by a finite difference expansion. Theoretically, the larger the number of sampling points, the closer the finite difference approaches the derivative. Practically, and due largely to the combined effects of local inhomogeneities in the conducting paper and the difficulty of exact probe placement, it is unprofitable to extend the finite difference sampling points beyond the minimum of two. The measurement of total flux now requires the following arrangements.

1. Measure the difference in potential between two sampling points astride the j-axis. Divide by the distance between the points to obtain the normal potential gradient.
2. Obtain the neighborhood current density by multiplying this gradient by the conductivity of the conducting paper.
3. Obtain the total normal current flowing through the neighborhood, by multiplying J by a, the diameter of the neighborhood.
4. Obtain the current flowing through the adjacent neighborhood by repetition of the same process.
5. Sum up the total current by adding the contributions of each neighborhood current. This gives the total flux between $\omega = 0$ and $\omega = \omega_p$. Within a constant of proportionality, this gives the phase shift.

The experimenter can do these things step by step. However, the process may and should be automated. The first four steps may easily be accomplished by means of two operational amplifiers, or one difference amplifier. It must be recognized, however, that errors will accumulate rapidly unless drift-free amplifiers are used for all subtractions. The use of chopper stabilized amplifiers is recommended.

The fifth step, the summation of the neighborhood currents suggests the use of an operational integrator. However, the independent variable in an operational integrator is unmistakeably time and the independent variable with respect to which the device must integrate is equally unmistakeably frequency. However, a change of variable is immediately effected by scanning the double row of probes along the frequency axis at a constant velocity. The variable ω then becomes a linear function of time, and the limits of integration change from $\omega = 0$ to $t = 0$ and $\omega = \omega_p$ to $t = t_p$. The display of phase shift vs frequency is conveniently plotted on a second beam of an oscilloscope. Both gain and phase may be synchronized by the use of a continuously rotating potentiometer which provides the horizontal sweep for both beams.

THEORY OF THE ROOT LOCUS PLOTTER

Since root loci are constituted of all s plane points at which Arg G(s) = $180^\circ + n360^\circ$, a branch of the root locus may be thought of as a path joining all the points whose phase shift with respect to the phase angle at zero frequency is precisely 180°. It enables one to obtain the closed loop response of a system from the open loop poles and zeros. As the previous discussion has shown, decibel gain and phase are conjugate harmonic functions of the complex frequency variable, s. Therefore loci of constant phase and loci of constant gain intersect orthogonally. In the analog, this means that loci of constant phse must intersect equipotentials at right angles. If, therefore, a point can be found in the s plane whose phase shift is 180° due to the allocation of the poles and zeros, the problem of finding a branch of the root locus is reduced to the problem of constraining a device to move in such a way as to intersect all equipotentials orthogonally.

A breakaway point on the real axis, for example, is a point of 180° phase. It also has the property that it is a point of minimum potential locally along the real axis, and is readily located in the model through this property. From a topological viewpoint, the breakaway point is a saddle point, and the root locus is a path of steepest descent.

It has been found practical to construct a small car which moves in such a manner as to cross all equipotential lines orthogonally. Two sensing probes connected to a difference amplifier detect the error in alignment of the drive sheels of the car, and by means of small servomotors drive the wheels in such a direction as to reduce this error to zero. The trace of the motion of the car is therefore a branch of the root locus.

The car itself is a closed loop system, and therefore need not have the luxury of chopper stabilized amplifiers or precision calibration. Since the device follows any line of constant phase, its use is not restricted to root loci, but may be extended to phase loci in general. Also, it is obvious that a 90° rotation of the pickup probes will constrain the car to follow any equipotential on which it is placed. The device may therefore be used for general purpose automatic field plotting. Also, since root locus techniques can be used to solve for roots of polynomials, the device may be readily adapted to this purpose.

THE INFINITE PLANE

In the foregoing discussion, it has been assumed that the model of the s plane is effectively infinite. Since any finite model will introduce errors at the boundary, it is necessary to either ignore these errors, compute and compensate for them, or transform a part of the plane in such a manner as to eliminate the finite-size error. Several schemes exist for this. A logarithmic transformation maps the upper half plane into a rectangular strip[1], but suffers the drawback that one must also "transform" his pole and zero locations, and also the resulting loci. Another well-known transformation maps the entire region exterior to a unit circle into another circle[7,8]. This transformation leaves the working region undistorted, and is to be preferred for some purposes. The unit circle "working region" and the transform region must be joined at each corresponding boundary point by a short circuiting path. However, these paths must be electrically isolated from each other. The double requirement of joining an infinite number of boundary points in one region to an infinite number in another region, without touching each other is theoretically formidable. In practice, however, the overall inhomogeneity and the local inhomogeneities of commercially available conducting paper render it unprofitable to make the spacing between connectors less than a few minutes of arc.

CONCLUSION

The practicability of the foregoing theory has been confirmed by the construction and testing of the described devices. The performance of gain and phase plotter was tested in a series of prototypes, by comparing the experimental results for a single complex pole pair with the gain and phase curves predicted by theory. The variation of the quality factor, Q, with pole position as well as change of phase at the half power points was quite satisfactory Other pole and zero configurations, including those for Chebyschev and Butterworth filters, as well as zero phase shift and constant gain functions gave experimental results which agreed well with theory.

The experimental root locus for three real axis poles also compared well with the theoretical curve for the same configuration and displayed the dependency of root locus on pole position satisfactorily. The prototypes already constructed seem to justify the claim that simple and easily constructed apparatus of this type has educational value in the undergraduate engineering laboratory.

The work done in constructing these devices was supported by a N.S.F. Science Teaching Equipment grant No. G22935.

REFERENCES

1. Scott, R. E., Network Synthesis by the Use of Potential Analogs. Proc. IRE, Vol. 40(7), 970-973 (1952).
2. Huggins, W. H., The Potential Analogue in Network Synthesis and Analysis, Air Force Cambridge Research Laboratories, E5066, March, 1951.
3. Darlington, S., The Potential Analogue Method of Network Synthesis, Bell System Tech. J., Vol. 30, 315-365 (1951).
4. Hansen, W. W., and Lundstrom, O. C., Experimental Determination of Impedance Functions by the Use of an Electrolytic Tank, Proc. IRE, Vol. 33(7), 528-534 (1945).
5. Boothroyd, A. R., and Westcott, J. H., The Application of the Electrolytic Tank to Servo-mechanism Design, Cranfield Conference, 1951.
6. Malavard, L., The Use of Rheolectrical Analogies in Aerodynamics, AGARDograph No. 18 (1956).
7. Boothroyd, A. R., Cherry, E. C. and Makar, R., An Electrolytic Tank for the Measurement of Steady State Response, Transient Response and other Allied Properties of Networks, J. IEE, Vol. 96, pt. 1, pp. 163-177 (1949).
8. Plonsey, R. and Collin, R. E., Principles and Applications of Electromagnetic Fields, McGraw Hill, New York, 1961, Chap. 5.
9. Van Valkenburg, M. E., Introduction to Modern Network Synthesis, John Wiley and Sons, New York, 1962, Chap. 9.
10. Guillemin, E. A., Theory of Linear Physical Systems, John Wiley and Sons, New York, 1963, Chap. 9.

FEEDBACK SYSTEM ANALYSIS BY FOUR-NODE GRAPHS

H. H. SUN
Electronic System Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts

and

G. C. REIS
Department of Electrical Engineering
Drexel Institute of Technology
Philadelphia 4, Pennsylvania

Synopsis

As an alternative to the study of feedback control system by the conventional single loop model, this paper presents the concept of the four-node graph (FNG). The FNG has a simple structure and allows multi-loop, multi-parameter calculations.

The FNG and the inverted FNG are first defined. From these graphs, expressions for return ratio, return difference, null return ratio and null return difference are developed. The FNG is then used to derive Blackman's relation in one variable. The generalized expression of Blackman's relation for n variables is derived by use of n interconnected FNG's which are subgraphs of the feedback system. Expressions are derived so that the terms of the generalized expression can be evaluated from a single FNG. A feedback transistor amplifier is used to exemplify the application of these results.

I. Introduction

A four-node flow graph (hereafter called FNG) is a signal flow graph which represents two simultaneous equations of two dependent variables expressed in terms of two independent variables. The FNG for a feedback system is constructed by use of return voltages (1, 2). In the study of feedback systems, it is often convenient to "open" the loop at a parameter and measure loop gains with various constraints on the input, output or other parameters of the system. This process defines the four nodes of the FNG; namely: the input node; the output node; plus the two nodes on both sides of the opened branch.

The usual analysis of feedback systems is based on a single-loop configuration having a forward gain of μ (or G) and a feedback gain of β (or H) (6). The FNG presents a more useful and flexible model for the study of feedback systems since it is more general than the single-loop structure, without an increase in complexity. The FNG can represent multi-loop structures, and readily accomodates investigation of the effect of parameter variations in these different loops. It also gives directly the necessary functions for the stability analysis of the system.

II FNG and the inverted FNG

The return ratio (T_k) of a feedback system with respect to a parameter k is the amount of signal returned to one side of the parameter k, when the loop is opened at k, and a unit signal is applied to the other side of k with the system input reduced to zero. This process is represented by figure (1), and the corresponding FNG is figure (2).

The transmittances of each of figure (2) are defined and expressed as follows:-

$$T_k = \frac{E_g}{E_b}\bigg|_{E_I = 0} = \frac{\Sigma_{(bg)}}{\Delta^{o}} = \text{return ratio} \tag{1}$$

$$G^{o} = \frac{E_o}{E_I}\bigg|_{E_b = 0} = \frac{\Sigma^{o}_{(I_o)}}{\Delta^{o}} = \text{direct transmission gain} \tag{2}$$

$$B = \frac{E_o}{E_b}\bigg|_{E_I = 0} = \frac{\Sigma_{(bo)}}{\Delta^{o}} = \text{backward fractionated gain} \tag{3}$$

$$C = \frac{E_g}{E_I}\bigg|_{E_b = 0} = \frac{\Sigma^{o}_{(Ig)}}{\Delta^{o}} = \text{forward fractionated gain} \tag{4}$$

where Δ is the determinant of the flow graph of the system as expressed by Mason (3, 4) and $\Sigma_{(ij)}$ is the numerator of Mason's general gain formula (3, 4) for the gain from node i to node j. The superscript "0" refers to the condition when the path k of the flow graph is broken to generate the FNG, as shown in Figure (1b) and figure (2).

Another graph which will be useful is the inverted FNG shown in figure (3). This is formed from the FNG by inversion (3, 4) of the direct transmission gain (G^{o}) of figure 2.

III Expressions for F_k and F_k' from FNG

The return differences F_k and null return differences F_k' are defined as follows:

$$F_k = \text{return difference} = 1 - T_k \tag{5}$$

$$F_k' = \text{null return difference} = 1 - T_k' \tag{6}$$

where T_k' is defined as the null return ratio which is the return ratio under the condition that the input signal is adjusted to give zero output signal. Sun and Reis (1) have shown that T_k' can be obtained from the FNG by inverting the path IO and finding the gain from node b to node g. This is shown as in Figure (3) or

$$T_k' = T_k - \frac{B\ C}{G^{o}} \tag{7}$$

It has also been shown (1) that

$$F_k = \frac{\Delta}{\Delta^{o}} \tag{8}$$

and

$$F_k' = \frac{\Sigma_{(IO)}}{\Sigma^{o}_{(IO)}} = \frac{\Sigma}{\Sigma^{o}} \tag{9}$$

Using this notation, the system gain (G) is:

$$G = \frac{E_o}{E_I} = \frac{\Sigma}{\Delta} \tag{10}$$

The gain (G_I) of the inverted graph (i.e. path IO inverted) is:

$$G_I = \frac{1}{G} = \frac{\sum_I}{\Delta_I} \tag{11}$$

where the subscript I indicates the inverted graph. Writing equation (11) as:

$$\frac{\sum_I}{\Delta_I} = \frac{\sum^o_I}{\Delta^o_I} \cdot \frac{\Delta^o_I}{\Delta_I} \cdot \frac{\sum_I}{\sum^o_I} \tag{12}$$

and noting

$$\frac{1}{G^o} = \frac{\sum^o_I}{\Delta^o_I} \tag{13}$$

Comparision of equations (12), (13) and (5) yields:

$$F_k = \frac{\sum_I}{\sum^o_I} \tag{14}$$

$$F_k{}' = \frac{\Delta_I}{\Delta^o_I} \tag{15}$$

Hence the return difference and null return differences can also be obtained from the use of inverted graph.

To summarize, it can be seen that F_k can be obtained by the use of (1) equations (1) and (5), (2) equation (8) directly, and (3) equation (14) of the inverted graph. Similarly $F_k{}'$ can be obtained from (1) equations (6) and (7) with the use of FNG, (2) equation (9) directly and (3) equation (15) of the inverted graph.

IV Blackman's equation (5) in one variable derived from the FNG

Reconnecting node b and node g of the FNG in figure (2) restores the system to its original condition as shown in figure (4). The gain of the system can thus be calculated from figure (4) as

$$G = G^0 \frac{F_k{}'}{F_k} \tag{16}$$

Equation (10) is the Blackman's equation in one variable.

V Blackman's equation in n variables by n FNG's.

Consider a system where two parameters, say k_1 and k_2 are of interest. The interconnected FNG's are shown in figure (5). Let $F_{k_1}(k_1, k_2)$ be the return difference with respect to k_1 when both k_1 & k_2 are in normal values. Let $F_{k_2}(0, k_2)$ be the return difference with respect to k_2, when k_2 is in normal value and $k_1 = 0$. Similar definitions apply to the null return differences. Let G^{00} represent the system gain when $k_1 = k_2 = 0$.

Using figure (6) and figure (8) we can obtain $F_{k_1}(k_1, k_2)$ by the use of equation (8), hence

$$F_{k_1}(k_1, k_2) = \frac{F_{k_{11}} \; F_{k_{22}} - T_{21}\, T_{12}}{F_{k_{22}}} \tag{17}$$

Similarly $F'_{k_1}(k_1, k_2)$ can be found by the use of equation (9), or

$$F'_{k_1}(k_1, k_2) = F_{k_{11}} - \frac{T_{21}\,T_{12} - \frac{B_1C_1}{G^{00}}\,F_{k_{22}} - \frac{B_1C_2}{G^{00}}\,T_{12} - \frac{B_2C_1}{G^{00}}\,T_{21}}{F_{k_{22}} + \frac{B_2C_2}{G^{00}}} \tag{18}$$

$F_{k_2}(0, k_2)$ and $F'_{k_2}(0, k_2)$ can be found from the FNG of figure (7), which is formed from figure (5) by letting $k_1 = 0$ or removing nodes b_1 and g_1. From figure (7), we have:

$$F_{k_2}(0, k_2) = F_{k_{22}} \tag{19}$$

and

$$F'_{k_2}(0, k_2) = F_{k_{22}} - \frac{B_2C_2}{G^{00}} \tag{20}$$

If nodes g_1 and b_1, and nodes g_2 and b_2, of figure (5) are reconnected as shown in figure (8), the system is restored to its original state, and the gain can be calculations:

$$G = G^{00}\left[1 + \frac{F_{k_{11}}\frac{B_2C_2}{G^{00}} + F_{k_{22}}\frac{B_1C_1}{G^{00}} + \frac{B_1C_2}{G^{00}}\,T_{12} + \frac{B_2C_1}{G^{00}}\,T_{21}}{F_{k_{11}}\,F_{k_{22}} - T_{12}\,T_{21}}\right] \tag{21}$$

which can be simplified to

$$G = G^{00}\,\frac{F'_{k_1}(k_1, k_2)}{F_k(k_1, k_2)}\,\frac{F'_{k_2}(0, k_2)}{F_{k_2}(0, k_2)} \tag{22}$$

From the above result for two parameters, the expression for n parameters can easily be derived as shown by figures (9), (10), and (11). The gain of figure (11) is thus seen to be

$$G = G^{00\ldots0}\,\frac{F'_{k_1}(k_1, k_2\ldots k_n)}{F_{k_1}(k_1\,k_2\ldots k_n)}\,\frac{F'_{k_2}(0, k_2\ldots k_n)}{F_{k_2}(0, k_2\ldots k_n)}\cdots\frac{F'_{k_n}(0, 0\ldots0, k_n)}{F_{k_n}(0, 0\ldots0, k_n)} \tag{23}$$

where $F_{k_i}(0, 0, \ldots 0, k_i, k_{i+1}, \ldots k_n)$ is the return difference with respect to k_i when all $k_j (j = 1, 2, \ldots i-1)$ have been reduced to zero and all other k's are in normal values. $F'_{k_i}(0, 0, \ldots 0, k_i, k_{i+1}, \ldots k_n)$ is similarly defined. $G^{00\ldots0}$ is the gain of the system with all the k's equal to zero. This result agrees with Blackman's generalized expression as derived by Mulligan (8).

VI Alternate Expression for Blackman's Equation in n-variables

Evaluating each of the n ratios of return differences in equation (23) involves gain calculations from a different FNG. In many cases it may be desirable to use only one FNG. Let $G^{(00\ldots0, k_i, k_{i+1}, \ldots k_n)}$ be the gain when all $k_j (j = 1, 2, \ldots i-1)$ are set to zero.

By use of equation (16) we can express:-

$$\frac{F'_{k_{n-1}}(0,\ 0,\ \ldots 0,\ k_{n-1},\ k_n)}{F_{k_{n-1}}(0,\ 0,\ \ldots 0,\ k_{n-1},\ k_n)} = \frac{G^{(0,0,\ldots 0,k_{n-1},\ k_n)}}{G^{(0,0,\ldots 0,k_n)}} \tag{24}$$

and

$$\frac{F'_{k_n}(0,0,\ldots 0,k_{n-1},\ k_n)}{F_{k_n}(0,0,\ldots 0,k_{n-1}\ k_n)} = \frac{G^{(0,0,\ldots 0,k_{n-1},\ k_n)}}{G^{(0,0,\ldots 0,k_{n-1},0)}} \tag{25}$$

Solving equation (25) for $G^{(0,0,\ldots 0,k_{n-1},\ k_n)}$ and substituting it into equation (24) yields

$$\frac{F'_{k_{n-1}}(0,0,\ldots 0,k_{n-1},k_n)}{F_{k_{n-1}}(0,0,\ldots 0,k_{n-1},k_n)} = \frac{G^{(0,0,\ldots k_{n-1},0)}}{G^{(0,0,\ldots 0,k_n)}} \quad \frac{F'_{k_n}(0,0,\ldots 0,k_{n-1},k_n)}{F_{k_n}(0,0,\ldots 0,k_{n-1},k_n} \tag{26}$$

Similarly it can be shown that

$$\frac{F'_{k_{n-2}}(0,0,\ldots 0,k_{n-2},k_{n-1},k_n)}{F_{k_{n-2}}(0,0,\ldots 0,k_{n-2},k_{n-1},k_n)} = \frac{G^{(0,0,\ldots 0,k_{n-2},k_{n-1},0)}}{G^{(0,0,\ldots 0,k_{n-1},k_n)}} \quad \frac{F'_{k_n}(0,0,\ldots 0,k_{n-2},k_{n-1},k_n)}{F_{k_n}(0,0,\ldots 0,k_{n-2},k_{n-1},k_n)} \tag{27}$$

or in general, we have:

$$\frac{F'_{k_i}(0,0,\ldots 0,k_i,k_{i+1},\ldots k_n)}{F_{k_i}(0,0,\ldots 0,k_i,k_{i+1},\ldots k_n)} =$$

$$\frac{G^{(0,0,\ldots 0,k_i,k_{i+1},\ldots k_{n-1},0)}}{G^{(0,0,\ldots 0,0,k_{i+1},\ldots k_n)}} \quad \frac{F'_{k_n}(0,0,\ldots 0,k_i,k_{i+1},\ldots k_n)}{F_{k_n}(0,0,\ldots 0,k_i,k_{i+1},\ldots k_n)} \tag{28}$$

Each of the terms of equation (23) can thus be evaluated by equation (28). The price paid for maintaining the same reference (say k_n) is that the resulting expression has 2 n quantities instead of n. However, all terms in the right hand side of equation (28) can be written by inspection from one single flow graph. Hence the result will be simpler to evaluate.

VIII Expression for Blackman's Equation in n Variables by one FNG and Its Inverse.

Substituting equations (9) and (14) into equation (23) yields:

$$G = G^{(0,0,\ldots 0)} \frac{\sum^{(k_1,k_2,\ldots k_n)}}{\sum^{(0,k_2,\ldots k_n)}} \frac{\sum^{(0,k_2,\ldots k_n)}}{\sum^{(0,0,\ldots k_n)}} \cdots$$

$$\frac{\sum_I^{(0,k_2\ldots k_n)}}{\sum_I^{(k_1,k_2\ldots k_n)}} \quad \frac{\sum_I^{(0,0,\ldots k_n)}}{\sum_I^{(0,k_2,\ldots k_n)}} \quad \ldots \tag{29}$$

$$G = G^{(0,0,\ldots 0)} \frac{\sum^{(k_1,k_2,\ldots k_n)}}{\sum^{(0,0,\ldots 0)}} \quad \frac{\sum_I^{(0,0,\ldots 0)}}{\sum_I^{(k_1,k_2,\ldots k_n)}} \tag{30}$$

$$G = G^{(0,0,\ldots 0)} \frac{F'_{nk}}{F_{nk}} \tag{31}$$

where: $$F'_{nk} = \frac{\sum^{(k_1,k_2,\ldots k_n)}}{\sum^{(0,0,\ldots 0)}} \tag{32}$$

$$F_{nk} = \frac{\sum_I^{(k_1,k_2,\ldots k_n)}}{\sum_I^{(0,0,\ldots 0)}} \tag{33}$$

The FNG of Figure (11) and its inverse are therefore the flow graphs required to obtain the Blackman's expression.

IX Sensitivity

Sensitivity (6) (S_k^G) of the gain G with respect to a parameter k is defined as:

$$S_k^G = \frac{k}{G}\,\frac{dG}{dk} = \frac{1}{F_k} - \frac{1}{F'_k} \tag{34}$$

Using the results obtained above, the following expressions for sensitivity can be readily derived.

$$S_k^G \doteq S_k^{F'_k} - S_k^{F_k} \tag{35}$$

$$S_k^G = S_k^{\Sigma} - S_k^{\Sigma_I} \tag{36}$$

$$S_k^G = \frac{\Sigma_I^o}{\Sigma_I} - \frac{\Sigma^o}{\Sigma} \tag{37}$$

$$S_k^G = \frac{\Sigma - \Sigma^o}{\Sigma} - \frac{\Sigma_I - \Sigma_I^o}{\Sigma_I} \tag{38}$$

Equation (38) is particularly well suited for calculating sensitivity for a large number of parameters. This is so since $\Sigma - \Sigma^o$ is simply all paths of Σ which contain k. Let these be called Σ'. (Similar remarks hold for Σ_I. Then equation (38) can be written

$$S_k^G = \frac{\Sigma'}{\Sigma} - \frac{\Sigma'_I}{\Sigma_I} \tag{39}$$

Once Σ and Σ_I have been determined, $S_{k_i}^G$ is found by selecting those paths of Σ and Σ_I which involve k_i. Substituting these into equation (39) for the various k_i yields the desired sensitivities.

X Example

To exemplify the use of the results presented in this paper consider the transistor amplifier of figure 12(a). The gain (G) of this amplifier (i_2/i_1) can be found in the following various manners using R_1 and R_L as variables.

$$G = G^{(R_1,\,0)} \frac{F'_{R_L}(R_1, R_L)}{F_{R_L}(R_1, R_L)} \tag{40}$$

$$= G^{(0,\,R_L)} \frac{F'_{R_1}(R_1\ R_L)}{F_{R_1}(R_1, R_L)} \tag{41}$$

$$= G^{(0,0)} \frac{F'_{R_1}(R_1, R_L)}{F_{R_1}(R_1, R_L)} \quad \frac{F'_{R_L}(0, R_L)}{F_{R_L}(0, R_L)} \tag{42}$$

$$= G^{(0,0)} \frac{F'_{R_L}(R_1, R_L)}{F_{R_L}(R_1, R_L)} \quad \frac{F'_{R_1}(R_1, 0)}{F_{R_1}(R_1, 0)} \tag{43}$$

$$= G^{00} \frac{\Sigma}{\Sigma_I} \quad \frac{\Sigma_I}{\Sigma}\Bigg|_{R_1 = R_L = 0} \tag{44}$$

From figure 12(b)

$$\Sigma = \beta - \frac{R_1}{R_f} = \frac{\beta R_f - R_1}{R_f} \tag{45}$$

From figure 12(c)

$$\Sigma_I = \frac{1}{\beta} + \frac{R_1}{\beta R_f} + \frac{R_L}{R_f} + \frac{R_L}{\beta R_f}$$

$$= \frac{R_f + R_1 + (\beta + 1) R_L}{\beta R_f} \tag{46}$$

By use of equations (8), (9), (14) or (15), the following are found:

$$F_{R_1}(R_1, R_L) = 1 + \frac{R_1}{R_f + (\beta + 1)R_L} \tag{47}$$

$$F'_{R_1}(R_1, R_L) = 1 - \frac{R_1}{\beta R_f} \tag{48}$$

$$F_{R_L}(R_1, R_L) = 1 + \frac{R_L(\beta + 1)}{R_1 + R_4} \tag{49}$$

$$F'_{R_L}(R_1, R_L) = F'_{R_L}(0, R_L) = 1 \tag{50}$$

$$F_{R_L}(0, R_L) = 1 + \frac{R_L(\beta + 1)}{R_f} \tag{51}$$

$$F_{R_1}(R_1, 0) = \frac{R_1 + R_f}{R_f} \tag{52}$$

$$F'_{R_1}(R_1, 0) = \frac{\beta R_f - R_1}{\beta R_f} \tag{53}$$

$$G^{(R_1, 0)} = \frac{\beta - R_1/R_f}{1 + R_1/R_f} \tag{54}$$

$$G^{(0, R_L)} = \frac{\beta}{1 + \frac{R_L}{R_f}(\beta + 1)} \tag{55}$$

$$G^{(0,0)} = \beta \tag{56}$$

Using the equations on the previous page to obtain the gain as shown in equations (40) to (45) we obtain the same result as:

$$G = \frac{\beta R_f - R_1}{R_f + R_1 + R_L(1 + \beta)} \tag{57}$$

Equation (57) can also be verified by use of Mason's Rule on figure 12(b). Substituting equations (45) and (46) into equations (39) or (37):

$$S^G_{R_1} = \frac{R_f + (\beta + 1) R_L}{R_1 + R_f + (\beta + 1) R_L} - \frac{\beta R_f}{\beta R_f - R_1} \tag{58}$$

$$S^G_{R_L} = \frac{R_1 + R_f}{R_1 + R_f + (\beta + 1) R_L} - 1$$

$$= -\frac{(\beta + 1) R_L}{R_1 + R_f + (\beta + 1) R_L} \tag{59}$$

XI Conclusion

The use of FNG has amply demonstrated by the derivation of the Blackman's expression, the expressions of the return differences, null return differences and sensitivity. The interconnected FNG enables one to derive these expressions in terms of n variables. These expressions are also expressed in terms of the determinant ($\triangle$) and the forward gain path ($\sum$) of the flow graph and its inverted flow graph. Therefore if the flow graph of a system has been obtained, the return differences, null return differences, sensitivity and the Blackman's expression can all be obtained, usually by inspection.

Bibliography

1. H. H. Sun and G. C. Reis - "General Methods of Finding the Null Return Difference From a Signal Flow Graph", Proceedings of the Sixth Midwest Symposium on Circuit Theory, May, 1963, pp. H1-H20.

2. G. C. Reis - "General Method for Determining Null Return Ratio", M.S. Thesis, Drexel Institute of Technology Library, June, 1962.

3. S. J. Mason - "Feedback Theory - Some Properties of Signal Flow Graphs", Proceedings IRE, Sept. 1953, pp. 1144-1156.

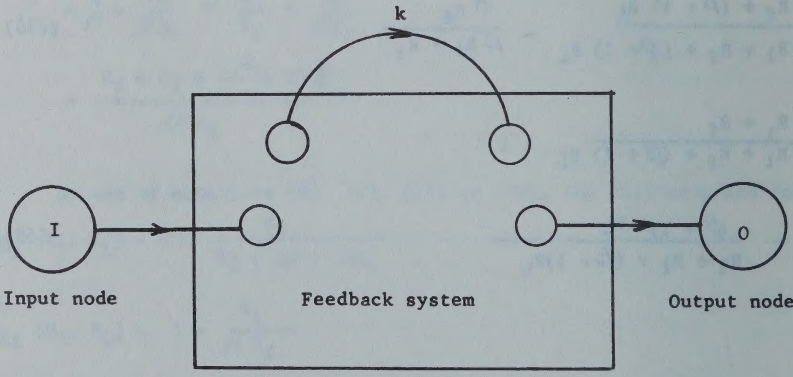
4. S. J. Mason - "Feedback Theory - Further Properties of Signal Flow Graphs", Proceedings IRE, July 1956, pp. 920-926.

5. R. B. Blackman - "Effect of Feedback on Impedance", Bell System Technical Journal, October, 1963, pp. 268-277.

6. H. W. Bode - "Network Analysis and Feedback Amplifier Design", Van Nostrand, 1945, pp. 44-102.

7. J. G. Truxal - "Control System Syntheses", McGraw-Hill, 1955, pp. 88-143.

8. J. F. Mulligan, Jr. - "Signal Transmission in Non-Reciprocal Systems", Proc. of the Symposium on Active Networks and Feedback Systems, July, 1960, pp. 125-153.



(1a)

Fig. (1a) Feedback system for single parameter "k".

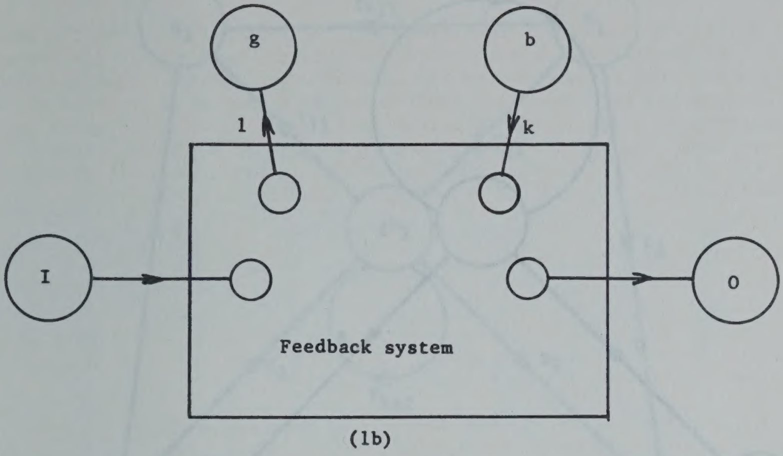

g
b
1
k
I
O
Feedback system
(1b)

Fig. (1b) Feedback system with path "k" broken

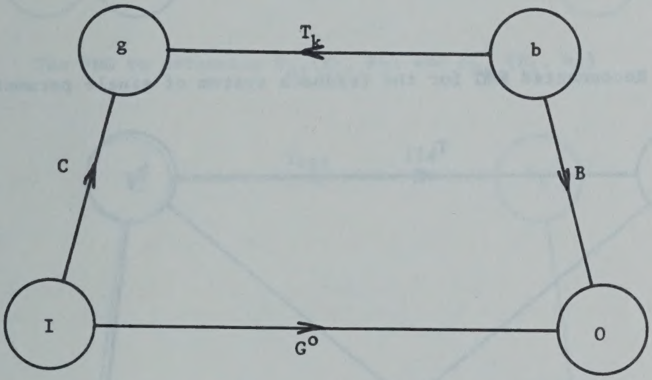

g
T_k
b
C
B
I
O
G^o

Fig. (2) FNG for the feedback system with single parameter.

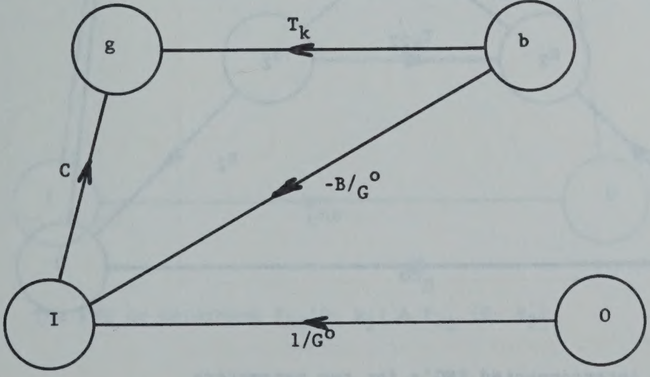

g
T_k
b
C
$-B/G^o$
I
O
$1/G^o$

Fig. (3) FNG with direct transmission gain path inverted.

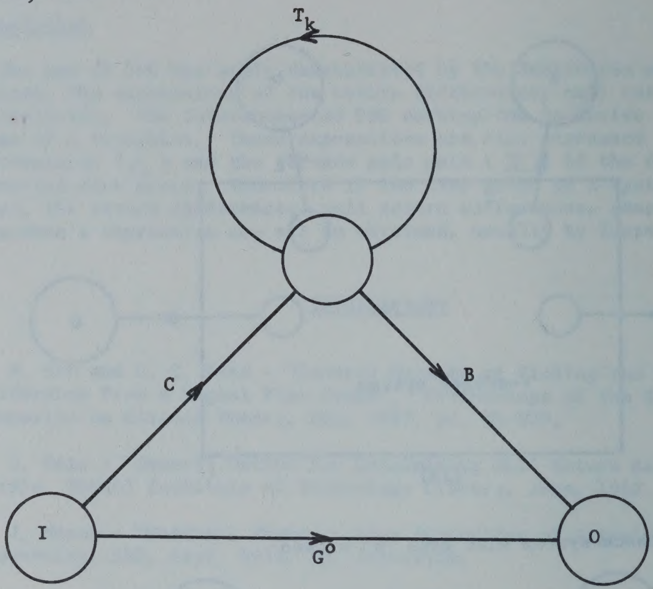


Fig. (4) Reconnected FNG for the feedback system of single parameter k.

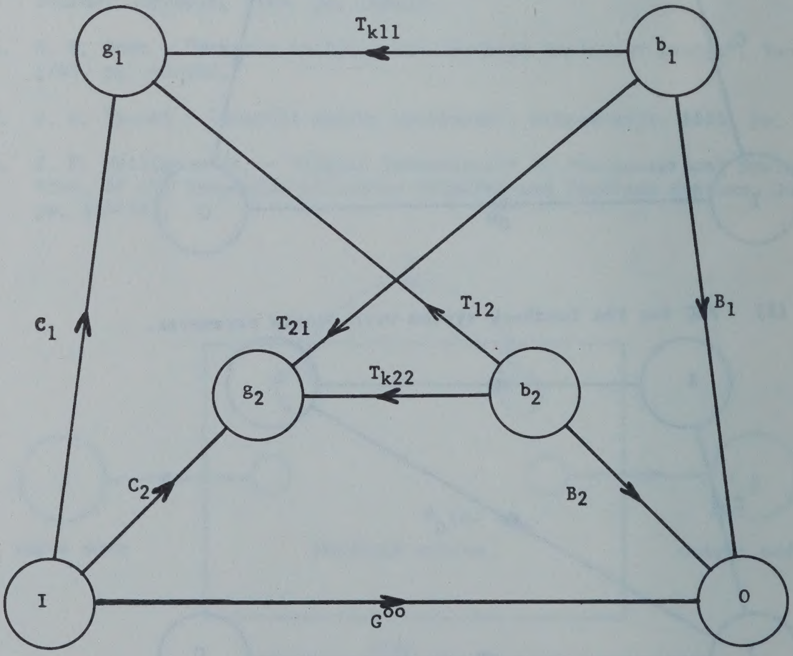

g_1
T_{k11}
b_1
C_1
T_{21}
T_{12}
B_1
g_2
T_{k22}
b_2
C_2
B_2
I
G^{oo}
O

Fig. (5) The interconnected FNG's for two parameters.

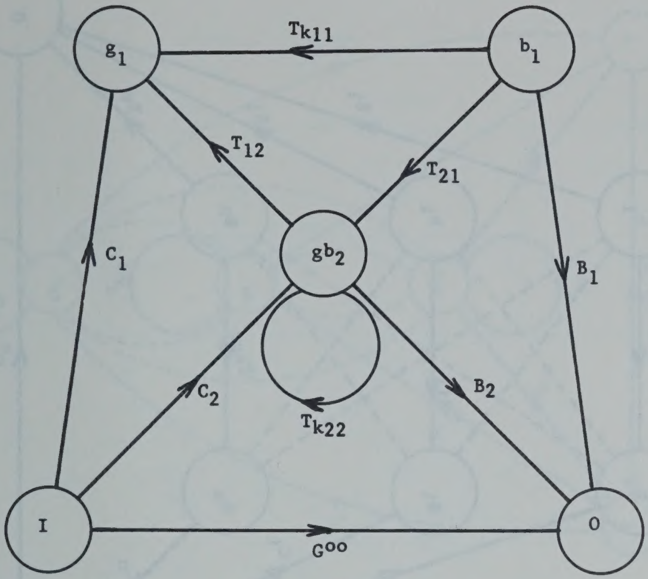

g_1
b_1
T_{k11}
T_{12}
T_{21}
gb_2
C_1
B_1
C_2
B_2
T_{k22}
I
O
G^{oo}

Fig. (6) The FNG to determine $F_{k1}(k_1, k_2)$ and $F'_{k1}(k_1, k_2)$

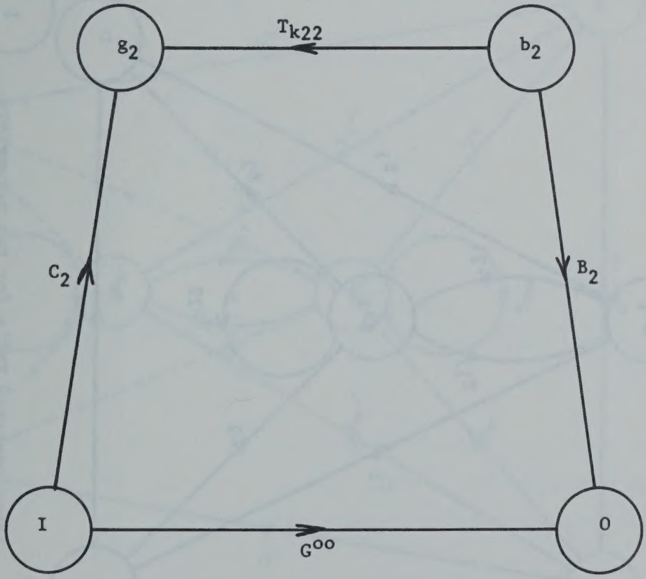

g_2
b_2
T_{k22}
C_2
B_2
I
O
G^{oo}

Fig. (7) The FNG to determine $F_{k2}(0, k_2)$ & $F'_{k2}(0, k_2)$

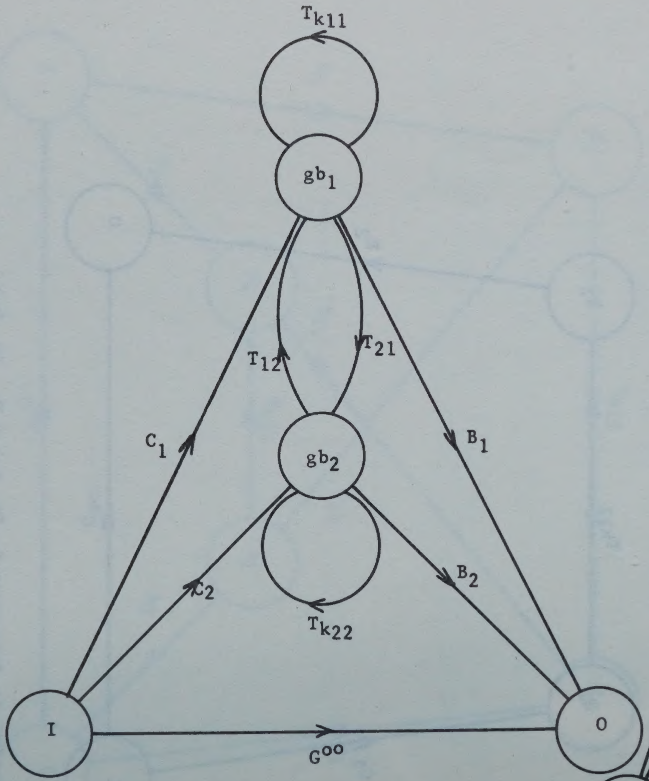

T_{k11}
gb_1
T_{12}
T_{21}
C_1
B_1
gb_2
C_2
B_2
T_{k22}
I
O
G^{oo}

Fig. (8) Interconnected FNG's for two parameters.

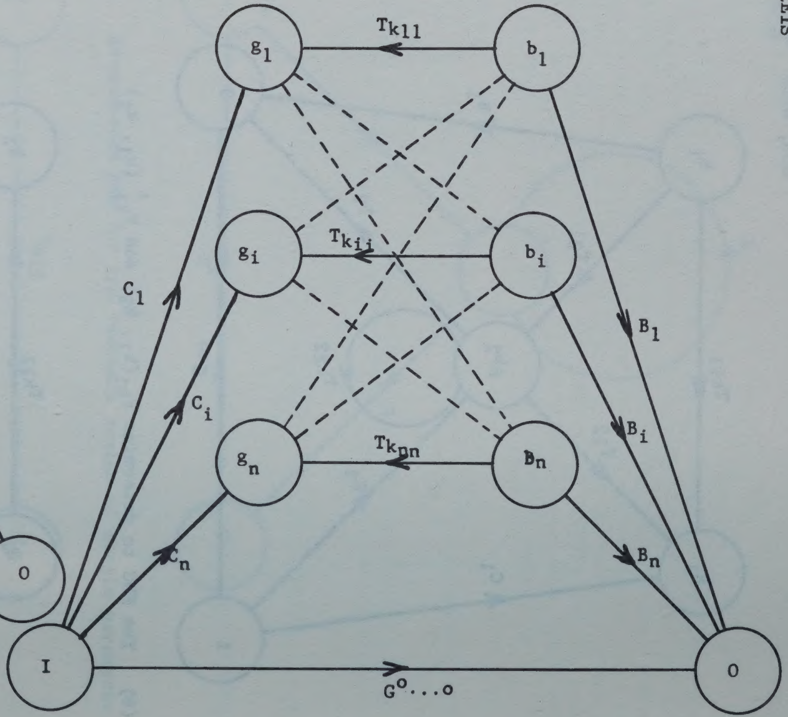

g_1
T_{k11}
b_1
g_i
T_{kii}
b_i
C_1
B_1
C_i
g_n
T_{knn}
b_n
B_i
C_n
B_n
I
O
$G^{o...o}$

Fig. (9) Interconnected FNG's with n parameters

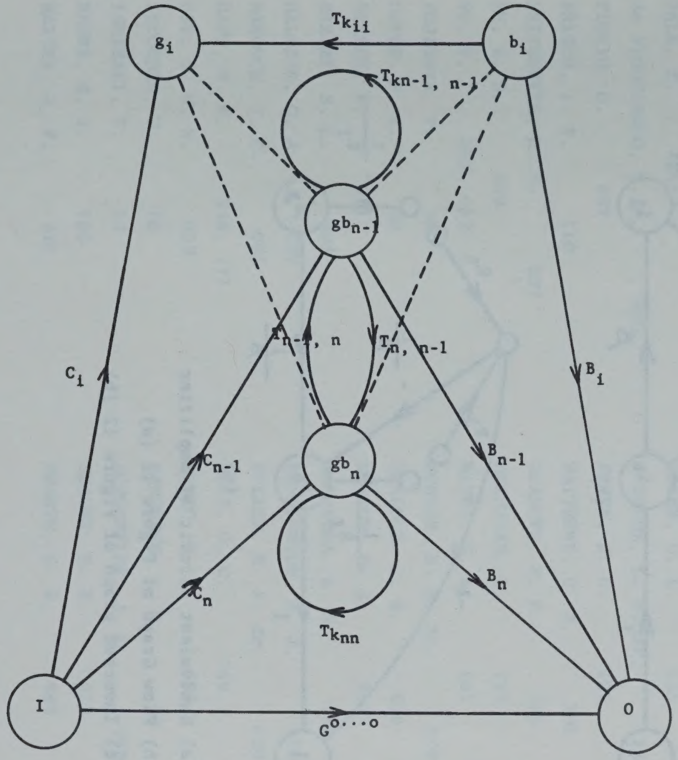

g_i
T_{k_ii}
b_i
T_{kn-1, n-1}
gb_{n-1}
T_{n-1, n}
T_{n, n-1}
C_i
B_i
C_{n-1}
B_{n-1}
gb_n
C_n
B_n
T_{knn}
I
O
G^{o...o}

Fig. (10) Interconnected FNG's for the evaluation $F_{k_i}(0, 0, \ldots 0, k_i, k_{i+1} \ldots k_n)$ and $F'_{k_i}(0,0,\ldots k_i, k_{i+1},\ldots k_n)$

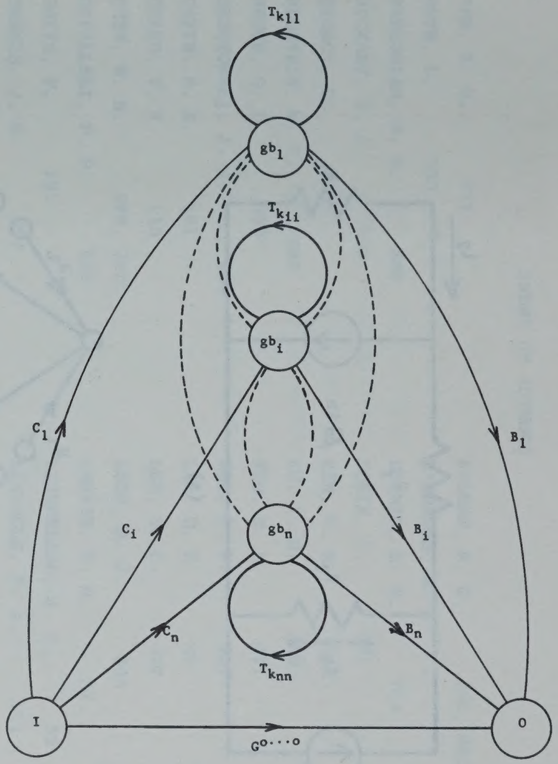

T_{k11}
gb_1
T_{kii}
gb_i
C_1
B_1
C_i
B_i
gb_n
C_n
B_n
T_{knn}
I
O
G^{o...o}

Fig. (11) Reconnected four-node flow graph for the feedback system with n parameters.

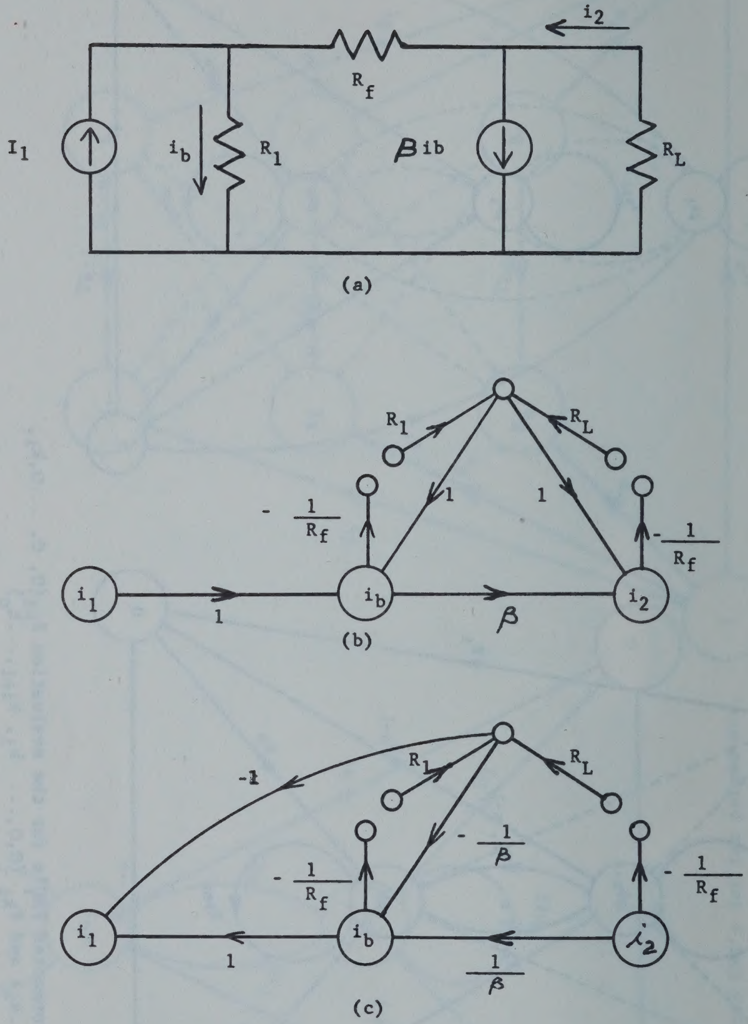

i2
Rf
I1
ib
R1
β ib
RL
(a)
R1
RL
1
1
- 1/Rf
- 1/Rf
i1
1
ib
β
i2
(b)
-1
R1
RL
- 1/β
- 1/Rf
- 1/Rf
i1
1
ib
1/β
i2
(c)

Figure 12 (a) Equivalent circuit of amplifier

(b) Flow Graph of Figure 12 (a)

(c) Inverted graph of Figure 12 (b)

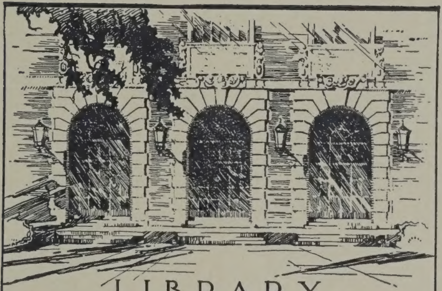
INDEX OF AUTHORS

ASH, R. B. 617
AUTH, L. 330
BEDROSIAN, S. D. 589
BICKART, T. A. 320
BIORCI, G. 537
BLOSTEIN, M. L. 458
BRUCE, G. D. 632
BRZOZOWSKI, J. A. 406
CAVIN, R. K. 191
CHANG, V. W. 116
CHEN, W. H. 275, 293
CIVALLERI, P. P. 537
DAVIS, F. 191
DENNIS, J. B. 367
DESOER, C. A. 50
EL-ABIAD, A. H. 205
FALK, T. 562
de FIGUEIREDO, R. P. 350
FIORIO, G. 537
FRISCH, I. T. 116
FRITZMEYER, R. D. 237
FU, K. S. 632
FU, Y. 521, 652
FUJISAWA, T. 385
GEFFE, P. R. 29
GERSCH, W. 621
HAKIMI, S. L. 394
HALIJAK, C. A. 237
HANCOCK, J. C. 661
HAPP, W. W. 148, 171
HSU, D. H. S. 652
IEDOKORO, T. 16
ISHIZAKI, Y. 16
KAHNE, S. J. 736
KAISER, J. F. 221
KALMAN, R. C. 456, 662
KARNAUGH, M. 1
KOENIG, H. E. 708
KOZIN, F. 621
KIM, W. H. 562
KU, Y. H. 309
KUH, E. S. 471
KUO, F. F. 427
LAI, D. C. 63
LEE, S. C. 427
LEON, B. J. 344
LERNER, R. M. 36
LINDENLAUB, J. C. 53
LINDHOLM, F. A. 171
LIU, B. 569
LIU, R. 330
LOSEE, D. L. 131
McBRIDE, L. E. Jr. 740
MCKEE, J. C. 191
MATTHEWS, G. A. 191
MEADOWS, H. E. 569
MEKSAWAN, T. 717
MITRA, S. K. 131
MORGAN, B. S. Jr. 509
MURDOCH, J. B. 675
MURPHY, G. J. 717
NARENDRA, K. S. 740
PENNINGTON, A. J. 579
PULLEN, K. A. Jr. 696
REIS, G. C. 764
REMEC, M. J. 502
SANDBERG, I. W. 99
SCHOEFFLER, J. D. 259
SEACAT, R. E. 613
SHARPE, G. E. 692

SIBBLES, G. M. 191
SUN, H. H. 764
SRINATH, M. D. 652
SZENTIRMAI, G. 37
TEMES, G. C. 246
THEON, B. 81
THOMASIAN, A. J. 50
TOKAD, Y. 708
TOTTY, R. E. 661
WATANABE, H. 16
WEINER, D. D. 81
WELSH, C. A. 757
YAU, S. S. 521
ZEMANIAN, A. H. 663

LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

[illegible]



LIBRARY
OF THE
UNIVERSITY
OF ILLINOIS

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

UNIVERSITY OF ILLINOIS-URBANA
621.31921AL54P C003
PROCEEDINGS.$URBANA ILL
1 1963
3 0112 006279555